

# A holistic deep learning framework for endoscopy

Ziang Xu

Green Templeton College

University of Oxford

*A thesis submitted for the degree of  
Doctor of Philosophy*

Michaelmas 2024

## Abstract

Endoscopy has developed into a routine method of examination and treatment for the digestive tract. Since most cancer patients will develop precancerous lesions in the early stages of cancer, for example, in colonoscopy, colon polyps are an early stage of colon cancer; in gastroscopy, lesions or ulcers in the mucosa of the stomach are an early sign of gastric cancer. Early endoscopy can detect early cancerous lesions before the onset of obvious symptoms, leading to earlier interventional treatment and reduced patient suffering. Endoscopy can help endoscopists to precisely locate the lesion area by visualising the internal structure of the organ and determine the nature of the lesion based on its shape, size and colour. However, diagnosis by clinicians is usually based on experience and training a specialised physician is time consuming. Therefore, the development of an automated and accurate vision algorithm for endoscopy can assist physicians in their examination to reduce the missed diagnosis of early cancer adenomas and precancerous lesions, thereby avoiding cancer prevention at an early stage and improving the survival rate of patients.

In this thesis, proposed holistic approaches tackle different challenges in endoscopy. Various tasks can assist navigation, treatment planing and minimise missed lesions. Specifically, a novel bounding box pruning approach is proposed to improve artefact detection and an image quality assessment system based on artefact detection is presented. This thesis also proposes new supervised learning and self-supervised learning based approaches for ulcerative colitis grading respectively. Meanwhile, the performance of the proposed self-supervised learning methods in different downstream tasks is explored. In addition, this thesis introduces a spatial-temporal self-attention mechanism with self-supervised learning as an auxiliary task to cope with the challenging problem of video polyp segmentation.

The original contributions are as follows: (1) Pruning-based deep learning artefact detection framework in endoscopy and quality assessment system. (2) Supervised ulcerative colitis grading algorithm using additive angular margin loss. (3) Self-supervised learning with composite pretext-class discrimination for improved generalisability in multi endoscopic image analysis tasks. (4) End-to-end self-supervised and temporal self-attention with feature branching for real-time video polyp segmentation.





# A holistic deep learning framework for endoscopy



Ziang Xu  
Green Templeton College  
University of Oxford

A thesis submitted for the degree of  
*Doctor of Philosophy*  
Michaelmas 2024

# Acknowledgements

First of all, I would like to express my deepest gratitude to my primary supervisor Prof. Jens Rittscher, for his invaluable support, unwavering trust, and continuous encouragement throughout my PhD journey. His insightful guidance has been instrumental in helping me explore the depth and breadth of my research interests, and I am deeply thankful for the freedom he provided me to pursue my own ideas. I am also immensely grateful to my co-supervisor Prof. Simon Leedham, for inviting me to watch endoscopic procedures and providing valuable comments from clinical medicine perspectives. I would also like to express my gratitude to my co-supervisor Dr. Sharib Ali, for his invaluable advice during my PhD study and guidance on publications. All of these have been invaluable in shaping this thesis.

I would like to extend my heartfelt thanks to the entire Rittscher research group. The stimulating discussions, collaborative spirit, and shared experiences have made this journey not only enriching but also enjoyable. I am fortunate to have worked alongside such talented and supportive colleagues.

I'm specially profoundly thankful to my parents for their unconditional love and enduring support, both emotionally and financially. Their belief in me has been a constant source of strength throughout this long journey, and I sincerely hope that I have met their hopes and expectations. I would also like to express my appreciation to my friends for their companionship, patience, and encouragement during this journey.

To everyone and everything I have met along the way, thank you for contributing to this significant chapter of my life. This journey may not have been smooth, but I still make it to the end. Finally, I want to say something to myself: The adventure of youth has come to an end, but your story is only just beginning.

# Abstract

Endoscopy has developed into a routine method of examination and treatment for the digestive tract. Since most cancer patients will develop precancerous lesions in the early stages of cancer, for example, in colonoscopy, colon polyps are an early stage of colon cancer; in gastroscopy, lesions or ulcers in the mucosa of the stomach are an early sign of gastric cancer. Early endoscopy can detect early cancerous lesions before the onset of obvious symptoms, leading to earlier interventional treatment and reduced patient suffering. Endoscopy can help endoscopists to precisely locate the lesion area by visualising the internal structure of the organ and determine the nature of the lesion based on its shape, size and colour. However, diagnosis by clinicians is usually based on experience and training a specialised physician is time consuming. Therefore, the development of an automated and accurate vision algorithm for endoscopy can assist physicians in their examination to reduce the missed diagnosis of early cancer adenomas and precancerous lesions, thereby avoiding cancer prevention at an early stage and improving the survival rate of patients.

In this thesis, proposed holistic approaches tackle different challenges in endoscopy. Various tasks can assist navigation, treatment planing and minimise missed lesions. Specifically, a novel bounding box pruning approach is proposed to improve artefact detection and an image quality assessment system based on artefact detection is presented. This thesis also proposes new supervised learning and self-supervised learning based approaches for ulcerative colitis grading respectively. Meanwhile, the performance of the proposed self-supervised learning methods in different downstream tasks is explored. In addition, this thesis introduces a spatial-temporal self-attention mechanism with self-supervised learning as an auxiliary task to cope with the challenging problem of video polyp segmentation.

The original contributions are as follows: (1) Pruning-based deep learning artefact detection framework in endoscopy and quality assessment system. (2) Supervised ulcerative colitis grading algorithm using additive angular margin loss. (3) Self-supervised learning with composite pretext-class discrimination for improved generalisability in multi endoscopic image analysis tasks. (4) End-to-end self-supervised and temporal self-attention with feature branching for real-time video polyp segmentation.

# Contents

|                                                             |             |
|-------------------------------------------------------------|-------------|
| <b>List of Figures</b>                                      | <b>viii</b> |
| <b>List of Abbreviations</b>                                | <b>xii</b>  |
| <b>1 Introduction</b>                                       | <b>1</b>    |
| 1.1 Motivation . . . . .                                    | 1           |
| 1.2 Thesis Structure and Contributions . . . . .            | 6           |
| 1.3 Publications . . . . .                                  | 8           |
| <b>2 Background</b>                                         | <b>10</b>   |
| 2.1 Clinical background . . . . .                           | 11          |
| 2.1.1 Gastrointestinal endoscopy . . . . .                  | 11          |
| 2.1.2 Gastrointestinal diseases monitoring . . . . .        | 12          |
| 2.1.2.1 Upper gastrointestinal disease monitoring . . . . . | 12          |
| 2.1.2.2 Lower gastrointestinal disease monitoring . . . . . | 13          |
| 2.2 Technical background . . . . .                          | 17          |
| 2.2.1 Image classification methods . . . . .                | 17          |
| 2.2.1.1 Traditional Methods . . . . .                       | 17          |
| 2.2.1.2 Deep learning-based Methods . . . . .               | 18          |
| 2.2.1.3 Medical image classification . . . . .              | 22          |
| 2.2.2 Image detection methods . . . . .                     | 25          |
| 2.2.2.1 Traditional Methods . . . . .                       | 25          |
| 2.2.2.2 Deep learning-based Methods . . . . .               | 26          |
| 2.2.2.3 Artefact detection . . . . .                        | 29          |
| 2.2.2.4 Polyp detection in endoscopy . . . . .              | 30          |
| 2.2.3 Image segmentation methods . . . . .                  | 31          |
| 2.2.3.1 Traditional Methods . . . . .                       | 31          |
| 2.2.3.2 Deep learning-based Methods . . . . .               | 32          |
| 2.2.3.3 SOTA segmentation methods in colonoscopy . . . . .  | 34          |
| 2.2.4 Relevant advances in deep learning . . . . .          | 36          |
| 2.2.4.1 Self-supervised learning . . . . .                  | 36          |
| 2.2.4.2 Attention mechanism . . . . .                       | 41          |
| 2.3 Summary . . . . .                                       | 44          |

|          |                                                                                                                |           |
|----------|----------------------------------------------------------------------------------------------------------------|-----------|
| <b>3</b> | <b>Computer-aided image quality assessment in endoscopy</b>                                                    | <b>46</b> |
| 3.1      | Introduction . . . . .                                                                                         | 47        |
| 3.2      | Methodology . . . . .                                                                                          | 48        |
| 3.2.1    | Artefact detection . . . . .                                                                                   | 48        |
| 3.2.2    | Novel pruning method using instance profiles . . . . .                                                         | 50        |
| 3.2.2.1  | Profile atlases . . . . .                                                                                      | 51        |
| 3.2.2.2  | Detection box pruning . . . . .                                                                                | 52        |
| 3.2.3    | Quality score assessment . . . . .                                                                             | 52        |
| 3.3      | Experiments . . . . .                                                                                          | 53        |
| 3.3.1    | Dataset . . . . .                                                                                              | 53        |
| 3.3.2    | Evaluation metrics . . . . .                                                                                   | 54        |
| 3.3.3    | Experimental setup . . . . .                                                                                   | 55        |
| 3.4      | Results . . . . .                                                                                              | 56        |
| 3.4.1    | Quantitative results . . . . .                                                                                 | 56        |
| 3.4.2    | Qualitative results . . . . .                                                                                  | 56        |
| 3.5      | Discussion & Conclusion . . . . .                                                                              | 58        |
| <b>4</b> | <b>Computer-aided risk assessment in ulcerative colitis</b>                                                    | <b>62</b> |
| 4.1      | Supervised additive angular margin loss and model scaling network<br>for optimised colitis scoring . . . . .   | 64        |
| 4.1.1    | Introduction . . . . .                                                                                         | 64        |
| 4.1.2    | Methodology . . . . .                                                                                          | 66        |
| 4.1.2.1  | Model scaling network . . . . .                                                                                | 67        |
| 4.1.2.2  | ArcFace loss . . . . .                                                                                         | 67        |
| 4.1.3    | Experiments . . . . .                                                                                          | 68        |
| 4.1.3.1  | Dataset . . . . .                                                                                              | 68        |
| 4.1.3.2  | Evaluation metrics and experimental setup . . . . .                                                            | 69        |
| 4.1.4    | Results . . . . .                                                                                              | 69        |
| 4.1.5    | Contribution and conclusion . . . . .                                                                          | 70        |
| 4.2      | Self-supervised learning with composite pretext-class discrimination<br>for improved colitis scoring . . . . . | 71        |
| 4.2.1    | Introduction . . . . .                                                                                         | 71        |
| 4.2.2    | Methodology . . . . .                                                                                          | 74        |
| 4.2.2.1  | Feature extraction (FE) block . . . . .                                                                        | 76        |
| 4.2.2.2  | Contrastive loss estimation with margin (CEM) block . . . . .                                                  | 77        |
| 4.2.2.3  | $k$ -means feature grouping . . . . .                                                                          | 79        |
| 4.2.2.4  | Cross-level discrimination at image and patch-levels . . . . .                                                 | 79        |
| 4.2.2.5  | Proposed loss . . . . .                                                                                        | 80        |
| 4.2.3    | Experiments . . . . .                                                                                          | 81        |

|          |                                                                                                                                   |           |
|----------|-----------------------------------------------------------------------------------------------------------------------------------|-----------|
| 4.2.3.1  | Dataset . . . . .                                                                                                                 | 81        |
| 4.2.3.2  | Evaluation metrics . . . . .                                                                                                      | 82        |
| 4.2.3.3  | Implementation details . . . . .                                                                                                  | 82        |
| 4.2.4    | Results . . . . .                                                                                                                 | 83        |
| 4.2.4.1  | Comparison for UC classification task . . . . .                                                                                   | 84        |
| 4.2.4.2  | Computation time . . . . .                                                                                                        | 84        |
| 4.2.4.3  | Generalisation . . . . .                                                                                                          | 84        |
| 4.2.4.4  | Ablation studies . . . . .                                                                                                        | 85        |
| 4.2.4.5  | Qualitative Analysis . . . . .                                                                                                    | 89        |
| 4.2.5    | Contribution and conclusion . . . . .                                                                                             | 91        |
| <b>5</b> | <b>Computer-aided colorectal polyp segmentation and detection</b>                                                                 | <b>95</b> |
| 5.1      | Motivation . . . . .                                                                                                              | 96        |
| 5.2      | Results from baseline and proposed self-supervised approach . . . .                                                               | 97        |
| 5.2.1    | Experiments . . . . .                                                                                                             | 97        |
| 5.2.1.1  | Baseline and Comparison Methods . . . . .                                                                                         | 97        |
| 5.2.1.2  | Dataset . . . . .                                                                                                                 | 97        |
| 5.2.1.3  | Evaluation metrics . . . . .                                                                                                      | 98        |
| 5.2.1.4  | Implementation details . . . . .                                                                                                  | 98        |
| 5.2.2    | Results . . . . .                                                                                                                 | 99        |
| 5.2.2.1  | Comparison for polyp detection task . . . . .                                                                                     | 99        |
| 5.2.2.2  | Comparison for polyp segmentation task . . . . .                                                                                  | 99        |
| 5.2.2.3  | Generalisability study for polyp segmentation . . . .                                                                             | 100       |
| 5.2.2.4  | Ablation studies . . . . .                                                                                                        | 101       |
| 5.2.2.5  | Qualitative Analysis . . . . .                                                                                                    | 103       |
| 5.2.3    | Contribution and conclusion . . . . .                                                                                             | 104       |
| 5.3      | End-to-end self-supervised and temporal self-attention with feature<br>branching for real-time video polyp segmentation . . . . . | 108       |
| 5.3.1    | Introduction . . . . .                                                                                                            | 108       |
| 5.3.2    | Methodology . . . . .                                                                                                             | 110       |
| 5.3.2.1  | Encoder self-supervision . . . . .                                                                                                | 111       |
| 5.3.2.2  | Normalised self-attention block . . . . .                                                                                         | 112       |
| 5.3.2.3  | Global-to-local feature learning . . . . .                                                                                        | 113       |
| 5.3.2.4  | Decoder . . . . .                                                                                                                 | 113       |
| 5.3.2.5  | Joint-loss optimisation (end-to-end framework) . . .                                                                              | 114       |
| 5.3.3    | Experimental setup . . . . .                                                                                                      | 114       |
| 5.3.3.1  | Datasets and Comparison models . . . . .                                                                                          | 114       |
| 5.3.3.2  | Evaluation metrics . . . . .                                                                                                      | 115       |
| 5.3.3.3  | Implementation details . . . . .                                                                                                  | 115       |

|          |                                                              |            |
|----------|--------------------------------------------------------------|------------|
| 5.3.4    | Results . . . . .                                            | 116        |
| 5.3.4.1  | Method comparisons . . . . .                                 | 116        |
| 5.3.4.2  | Generalization . . . . .                                     | 118        |
| 5.3.4.3  | Ablation Study . . . . .                                     | 120        |
| 5.3.4.4  | Qualitative results . . . . .                                | 122        |
| 5.3.5    | Discussion . . . . .                                         | 122        |
| 5.3.6    | Conclusion . . . . .                                         | 125        |
| <b>6</b> | <b>Conclusion and Future work</b>                            | <b>126</b> |
| 6.1      | Summary of Contributions . . . . .                           | 126        |
| 6.2      | Limitations and directions for future work . . . . .         | 128        |
| 6.2.1    | Image quality enhancement based on artefact restoration . .  | 128        |
| 6.2.2    | More comprehensive grading strategy for ulcerative colitis . | 129        |
| 6.2.3    | Quantification of disease burden . . . . .                   | 130        |
| 6.3      | Conclusion . . . . .                                         | 131        |
|          | <b>References</b>                                            | <b>133</b> |

# List of Figures

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |    |
|-----|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1.1 | Images on the right shows two classic diseases, rectal polyps and ulcerative colitis. . . . .                                                                                                                                                                                                                                                                                                                                                                                   | 3  |
| 1.2 | Samples of various image-level annotated categories in the upper and lower gastrointestinal tract. The image intercepts from the original paper. [3] . . . . .                                                                                                                                                                                                                                                                                                                  | 5  |
| 1.3 | Overview of the thesis framework. The left column is the normal manual dignosis. Proposed AI assisied diagnosis framework is on the right column. . . . .                                                                                                                                                                                                                                                                                                                       | 7  |
| 2.1 | The endoscope is mainly composed of 4 major structural components, including control body, insertion tube, bending section/distal end and light guide connector [4]. Mechanical system, plumbing system and illumination system are respectively connected with various components, which provides a variety of functions for the endoscope.                                                                                                                                    | 12 |
| 2.2 | Samples of upper gastrointestinal diseases [11]. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                        | 13 |
| 2.3 | Images shows two classic diseases. The first line presents polyps and the second line presents 3 classes ulcerative colitis scored by Mayo endoscopic scoring. . . . .                                                                                                                                                                                                                                                                                                          | 14 |
| 2.4 | Architecture of LetNet-5 [29] and AlexNet [30]. The images intercept from the original paper. . . . .                                                                                                                                                                                                                                                                                                                                                                           | 19 |
| 2.5 | Architecture of VGGNet [33], GoogLeNet [34] and inception module. VGGNet and GoogLeNet utilise deeper network layers, in which Inception Module in GoogLeNet can not only maintain the sparsity of the network structure, but also perform high-performance calculations in dense matrices. The images intercept from the original paper. . .                                                                                                                                   | 20 |
| 2.6 | Architecture of ResNet [35], DenseNet [36] and model scaling in EfficientNet [37]. The residual learning in ResNet effectively solves the gradient vanishing problem in model training through identity mapping. DenseNet utilises the dense connection mechanism to allow the network to receive more information, while EfficientNet uses the Neural Architecture Search method to search for the best model structure. The images intercept from the original paper. . . . . | 22 |



|      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |    |
|------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.7  | Architecture of YOLO-v1 [62]. The image intercepts from the original paper. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | 26 |
| 2.8  | Architecture of FCN [95] and U-net [96]. The images intercept from the original paper. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | 33 |
| 2.9  | Typical paradigm in contrastive learning of visual representations [121]. The original input image $x$ undergoes two different data augmentations ( $t \sim T$ and $t' \sim T$ ) in the same series to generate two related views ( $\tilde{x}_i$ and $\tilde{x}_j$ ). $f(\cdot)$ and $g(\cdot)$ are the encoder network and projection head, respectively. The agreement between feature embeddings ( $z_i$ and $z_j$ ) is increased by minimizing the contrastive loss during the training phase. The representation $h$ is used for downstream tasks. . . . .           | 38 |
| 2.10 | Architecture of generative self-supervised learning [126]. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | 38 |
| 2.11 | Architecture of BYOL [129]. Online network and target network have similar structure with encoder $f$ and projector $g$ . The difference is that the online network updates parameters $\theta$ through training, and the parameters $\xi$ of the target network are an exponential moving average of the online parameters. The mean squared error is minimised between the online network output through the predictor $q_\theta(z_\theta)$ and the target network output through the stop-gradient $sg(z'_\xi)$ . The image intercepts from the original paper. . . . . | 39 |
| 2.12 | Architecture of BEiT [132]. The strategy for pre-training involves reconstructing the original visual tokens from the corrupted image patches. The image intercepts from the original paper. . . . .                                                                                                                                                                                                                                                                                                                                                                       | 40 |
| 2.13 | Architecture of Spatial Transformer Network [135]. The image intercepts from the original paper. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | 41 |
| 2.14 | Architecture of Vision Transforme [131]. The image intercepts from the original paper. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | 43 |
| 2.15 | Architecture of Scaled Dot-Product Attention and Multi-Head Attention [139]. The image intercepts from the original paper. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                         | 44 |
| 3.1  | The architecture of proposed method. The overall process is divided into two parts, the artefact detection network and post-processing block based on artefactinstance profiles. . . . .                                                                                                                                                                                                                                                                                                                                                                                   | 49 |
| 3.2  | The mean ( $\mu$ ) and standard deviation ( $\sigma$ ) for center-line of 7 different artefact classes created from ground truth samples. The blue bars represent the average greyscale value for each artefact category. The red bars represent the corresponding standard deviation, indicating the spread of greyscale distribution within each category. . . . .                                                                                                                                                                                                       | 50 |
| 3.3  | Examples of artefactlocations in the ground truth image. The size of bounding boxes belonging to artefactclass (green) are analogous and they are all on or close to the center line of bounding boxes. . . . .                                                                                                                                                                                                                                                                                                                                                            | 51 |

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |    |
|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 3.4 | Location weight distribution in the quality scoring assessment. . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               | 54 |
| 3.5 | Qualitative Results of EfficientDet-d0, EfficientDet-NMS and EfficientDet-NMS-PDF. Different colored bounding boxes represent different classes.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    | 57 |
| 3.6 | <i>QS</i> assessment focus on artefact detection, location and area [140]. The left image shows the image with detection boxes in the artefact detection task. The right image shows the corresponding area percentage. On the far right are their quality scores. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | 59 |
| 4.1 | Mayo endoscopic scoring representative images . . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | 65 |
| 4.2 | Training of a model scaling network for improved colitis classification with ArcFace loss [149]. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 65 |
| 4.3 | Endoscopic image analysis for ulcerative colitis scoring. (on left) Representative images for Mayo endoscopic scoring from 0 to 3, and (on right) t-SNE plots [168] with perplexity of 15 and maximum iterations of 1000 for all test samples before learning and after supervised learning using ResNet50 [35]. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 73 |
| 4.4 | Block diagram of our proposed self-supervised learning framework with composite pretext-class discrimination losses. The ResNet50 encoder network [35] is fine-tuned using original images with transformations and randomly shuffled patches, simulating a jigsaw puzzle task as in [120], within a self-supervised setting. This setup encourages the model to learn spatially invariant and discriminative features, improving generalisability and performance in downstream tasks. A contrastive estimator with margin is separately computed for both image-level and patch-level instances. Additionally, a group-wise contrastive loss is introduced by comparing the centroids at the patch level ( $PC^k$ ) and the image level ( $C^k$ ). A memory bank $\mathcal{M}$ [169] is utilised to store all feature representations for efficient contrastive learning. . . . . | 75 |
| 4.5 | a) t-SNE plot for the raw test data, baseline network (supervised) and two SSL approaches with perplexity of 5 in a. i) and 15 in a. ii); and b) attention maps of the proposed approach compared to other state-of-the-art methods for multi-class ulcerative colitis classification task (MES 0, MES 1, and MES 2). Principal Components Analysis was used to reduce the dimension of the feature maps to 50 as suggested in [168]. It is to be noted that t-SNE plots were obtained before 1000 iterations (set maximum) as the t-SNE converged, i.e., the same error after a certain number of iterations. The minimisation cost (Kullback-Leibler divergence) error was higher for all plots for the perplexity of 5 compared to the perplexity of 15. . . . .                                                                                                                 | 90 |

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |     |
|-----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 4.6 | Limitations of our proposed approach for different data conditions, including blur, difficult scene views due to organ topology or camera positions, imaging artefacts such as specularities and pixel saturation.                                                                                                                                                                                                                                                                                              | 92  |
| 5.1 | Qualitative comparison of our proposed self-supervised approach with other state-of-the-art methods for polyp detection tasks. . . . .                                                                                                                                                                                                                                                                                                                                                                          | 105 |
| 5.2 | Qualitative comparison of our proposed self-supervised approach with other state-of-the-art methods for polyp segmentation task. . .                                                                                                                                                                                                                                                                                                                                                                            | 106 |
| 5.3 | Limitations of our proposed approach in polyp detection and segmentation tasks. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                         | 107 |
| 5.4 | Proposed network: Our network consists of three parts: a) Global and local encoders trained with a self-supervision loss comprising Jigsaw-puzzle sampling. b) High-level global and local features branch from both encoders and pass through a normalised self-attention block [116], enabling global-to-local feature learning. c) A decoder layer that fuses the self-attention high-level feature maps from NS blocks and the low-level features from the local encoder for final mask prediction. . . . . | 109 |
| 5.5 | Normalised self-attention block used in our proposed network. . . .                                                                                                                                                                                                                                                                                                                                                                                                                                             | 110 |
| 5.6 | Qualitative results: Easy samples from SUN-SEG-Easy (unseen). . .                                                                                                                                                                                                                                                                                                                                                                                                                                               | 121 |
| 5.7 | Qualitative results: Hard samples from SUN-SEG-Hard (unseen, top rows), and samples from unseen data centre CVC-612 (bottom rows).                                                                                                                                                                                                                                                                                                                                                                              | 123 |
| 5.8 | Limitations of our proposed approach, including bubbles, instruments, and imaging artefacts such as specularities and pixel saturation. . . .                                                                                                                                                                                                                                                                                                                                                                   | 124 |

# List of Abbreviations

|                 |                                                               |
|-----------------|---------------------------------------------------------------|
| <b>UC</b>       | Ulcerative Colitis                                            |
| <b>AI</b>       | Artificial Intelligence                                       |
| <b>CT</b>       | Computed Tomography                                           |
| <b>MRI</b>      | Magnetic Resonance Imaging                                    |
| <b>IBD</b>      | Inflammatory Bowel Disease                                    |
| <b>MCES</b>     | Mayo Clinic Endoscopic Subscore                               |
| <b>CD</b>       | Crohn's Disease                                               |
| <b>NDBE</b>     | Nondysplastic Barrett's Esophagus                             |
| <b>EAC</b>      | Esophageal Adenocarcinoma                                     |
| <b>CRC</b>      | Colorectal Cancer                                             |
| <b>GLCM</b>     | Gray-Level Co-occurrence Matrix                               |
| <b>LBP</b>      | Local Binary Pattern                                          |
| <b>PCA</b>      | Principal Component Analysis                                  |
| <b>LDA</b>      | Linear Discriminant Analysis                                  |
| <b>SVM</b>      | Support Vector Machine.                                       |
| <b>K-NN</b>     | K-nearest Neighbor Algorithm                                  |
| <b>CNN</b>      | Convolutional Neural Network                                  |
| <b>ReLU</b>     | Rectified Linear Unit                                         |
| <b>ResNet</b>   | Deep Residual Neural Network                                  |
| <b>DenseNet</b> | Densely Connected Neural Network                              |
| <b>FCN</b>      | Fully Convolutional Neural Network                            |
| <b>mDice</b>    | Mean Dice Similarity Coefficient                              |
| <b>HOG</b>      | Histogram of Directed Gradients                               |
| <b>YOLO</b>     | You Only Look Once (a one-stage region-free detection method) |
| <b>FPN</b>      | Feature Pyramid Net                                           |

|              |           |                                           |
|--------------|-----------|-------------------------------------------|
| <b>R-CNN</b> | . . . . . | Region-based Convolutional Neural Network |
| <b>SSL</b>   | . . . . . | Self-supervised Learning                  |
| <b>PIRL</b>  | . . . . . | Pretext Invariant Representation Learning |
| <b>ViT</b>   | . . . . . | Vision Transformer.                       |
| <b>mAP</b>   | . . . . . | mean Average Precision                    |
| <b>IoU</b>   | . . . . . | Intersection-over-Union                   |
| <b>PDF</b>   | . . . . . | Probability Density Function              |
| <b>NMS</b>   | . . . . . | Non-Maximum Suppression                   |
| <b>QWK</b>   | . . . . . | Quadratic Weighted Kappa                  |
| <b>DSC</b>   | . . . . . | Dice Similarity Coefficient               |
| <b>NCE</b>   | . . . . . | Noise Contrastive Estimator               |
| <b>SOTA</b>  | . . . . . | State-of-the-art                          |



# 1

## Introduction

### Contents

---

|            |                                                     |          |
|------------|-----------------------------------------------------|----------|
| <b>1.1</b> | <b>Motivation . . . . .</b>                         | <b>1</b> |
| <b>1.2</b> | <b>Thesis Structure and Contributions . . . . .</b> | <b>6</b> |
| <b>1.3</b> | <b>Publications . . . . .</b>                       | <b>8</b> |

---

### 1.1 Motivation

The human gastrointestinal system is very fragile and vulnerable to viruses. It is estimated that gastrointestinal diseases, including gastric cancer and colorectal cancer, cause about 1.8 million deaths each year, accounting for more than 12% of the global mortality rate [1]. Disruptions to intestinal homeostasis, including microbial imbalances and immune dysregulation, can lead to inflammation and various rectal diseases such as rectal polyps, ulcerative colitis (UC), and intestinal adenomas. These conditions can cause significant damage to the rectal mucosa, increasing the risk of disease progression and potentially leading to colorectal cancer.

Endoscopy is a widely used clinical procedure for the early screening of cancer. Endoscopes provide clinicians with a way to directly observe the internal conditions of the body through a camera that enters the body and simultaneously transmits video images to an external monitor. It is recommended that patients undergo

regular endoscopic monitoring and generate a disease evolution map to track the development of the disease. Early detection of lesions allows for earlier treatment. As a clinical procedure, endoscopy plays a pivotal role in improving the detection rate of gastrointestinal diseases by allowing direct visualisation of the mucosal surface. However, the accuracy of current endoscopy relies heavily on the operation of professional doctors. The doctor must continuously rotate the endoscope or adjust the patient’s body so that the endoscope can cover most of the area. But relying on local observation and evaluation based on clinician’s experience is inefficient — especially in polyposis syndrome, where polyps are numerous and distributed throughout most of the gastrointestinal tract. Relying only on clinicians’ experience for judgement is prone to error, which can affect the detection rate of cancer, precancerous lesions, and inflammation. At the same time, during endoscopic surgery, due to the complex gastrointestinal environment, lighting conditions, and unstable motion, a large number of artefacts and occlusions inevitably appear in the video, affecting both image quality and the clinician’s diagnosis.

In recent years, the application of Artificial Intelligence (AI), particularly deep learning algorithms, has significantly expanded within the realm of computer vision, yielding commendable results. Within the domain of medical image analysis, AI has demonstrated its capacity to augment clinical diagnostics through its exceptional performance. This includes the automated detection of lung nodules in computed tomography (CT) scans and the identification of cardiac anomalies in magnetic resonance imaging (MRI), etc. Furthermore, in the context of endoscopic procedures such as colonoscopy and gastroscopy, AI has the capability to autonomously recognise and analyse abnormalities like colon polyps and gastric carcinoma. This aids medical professionals in pinpointing areas of concern, thereby substantially reducing the incidence of overlooked diagnoses.

However, there is still a lack of a complete AI-assisted system to support clinicians in endoscopic examinations. This thesis proposes a holistic colonoscopy-assisted detection framework that includes multiple functions to meet clinical needs, and



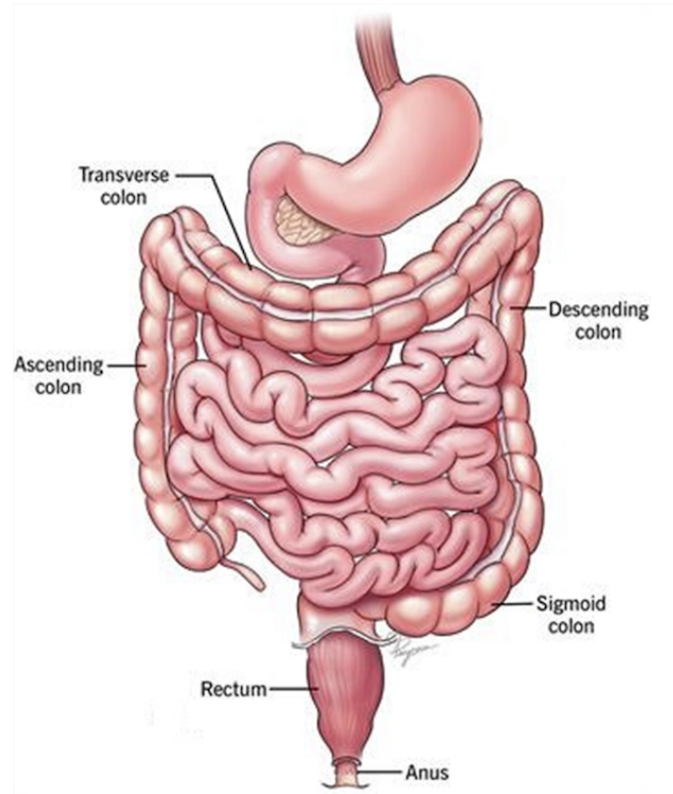


**Figure 1.1:** The image shows the biological interpretation of colon, including the ascending, transverse, descending and sigmoid colon [2].

introduces deep learning-based approaches to address the challenges in endoscopic image analysis. These challenges and corresponding solutions are as follows:

- **Image quality assessment:** Image quality is an important indicator for evaluating colonoscopy. In addition to improving image quality from the hardware perspective, such as using a higher-resolution camera, the operation of clinicians during colonoscopy also greatly affects image quality. Incorrect operation leads to lower image quality, which may affect clinicians' judgement and increase surgical risks.

**Solution:** The image quality of each frame in the colonoscopy video can be effectively evaluated by automatically detecting seven common artefacts, including saturation, specularities, bubbles, contrast, blood, blur, and imaging artefacts. This thesis presents an automatic artefact detection framework based on deep learning, which is further refined through the pruning of instance profiles to enhance detection accuracy. Each artefact detection frame is scored

and evaluated, facilitating the determination of an overall image quality score for each frame.

- **Automatic and precise disease detection:** The thesis focuses on two diseases in endoscopy: ulcerative colitis (UC) and polyps. UC, a typical representative of inflammatory bowel disease (IBD), can cause serious damage to the rectal mucosa, leading to rectal diseases and even cancer. Similarly, polyps are also an early manifestation of cancer. Early identification and prompt intervention are crucial for mitigating the risk of cancer development. However, in endoscopy, the endoscopist’s judgement of the lesion on the screen is made by the naked eye, which is affected by a number of factors that result in a high rate of under-detection of smaller polyps. Similarly, the evaluation of UC severity is predominantly reliant on the experience of clinicians, which may vary significantly.

**Solution:** The implementation of an automated and precise grading system for UC can substantially support clinicians in their diagnostic assessments and diminish discrepancies observed between colonoscopy evaluations and UC follow-up assessments. The detection and segmentation of polyps can more accurately quantify lesions and provide support for subsequent surgery. This thesis proposes a deep learning approach grounded in supervised learning that leverages additive angular margin loss to enhance the accuracy of multi-class UC grading, specifically utilising the Mayo Endoscopic Score (MES) as a scoring criterion. Meanwhile, a self-attention framework applied to video polyp segmentation is proposed in Chapter 5.

- **Scarcity of annotated data:** Although a large amount of video data can be generated in endoscopic surgery, the lack of annotated data for supervised learning still presents a significant challenge. The creation of standard medically labelled datasets necessitates extensive manual annotation by professional clinicians for each video frame, a process that is both time-consuming and labour-intensive. Moreover, the variability in annotations

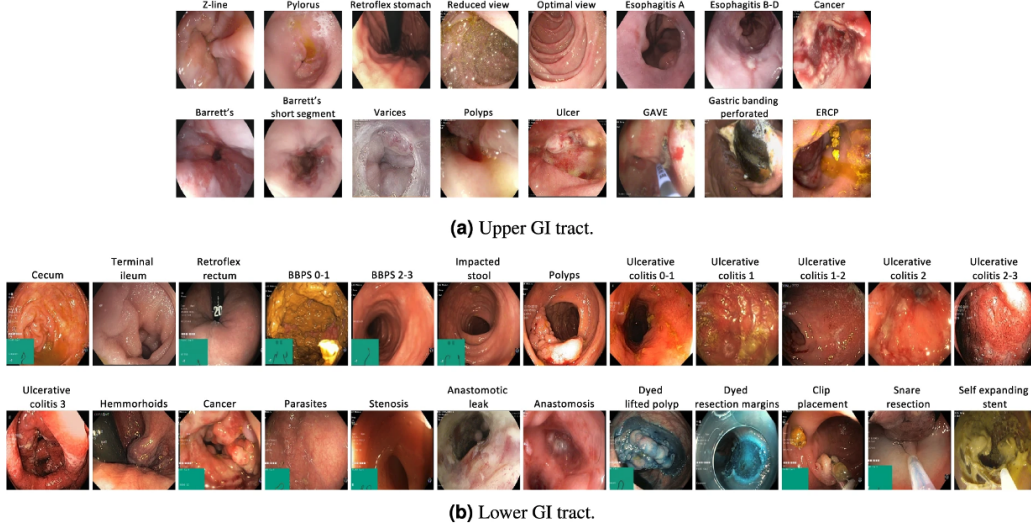


**Figure 1.2:** Samples of various image-level annotated categories in the upper and lower gastrointestinal tract. The image intercepts from the original paper. [3]

among different clinicians further complicates this issue.

**Solution:** The adoption of self-supervised learning techniques offers a promising solution to this challenge. Self-supervised learning enables the development of a pre-trained model through training on raw, unlabelled data, which can subsequently be fine-tuned for specific downstream tasks, achieving state-of-the-art (SOTA) outcomes. This thesis introduces a novel self-supervised learning framework that amalgamates clustering and contrastive learning strategies, which is capable of extracting semantically meaningful features from a vast corpus of unlabelled data. The proposed method can be applied to various downstream tasks pertinent to endoscopic analysis, such as polyp segmentation, polyp detection, and the grading of ulcerative colitis.

This thesis proposes a deep learning-based method that can solve the above problems and provide feedback to assist clinicians in diagnosis. Fig 1.2 shows some samples of anatomical landmarks as well as pathological and normal findings in the HyperKvasir dataset [3]. As can be seen in Fig. 1.2, image analysis based on endoscopy datasets is challenging. For instance, boundaries of certain polyps are difficult to distinguish from the color of the colon, and UC with varying

MES are visually similar. The core technical contributions of this thesis are summarized as follows:

- A pruning-based deep learning artefact detection framework in endoscopy.
- An end-to-end supervised method for automatic grading ulcerative colitis.
- Self-supervised learning with composite pretext-class discrimination for improved generalisability in endoscopic image analysis.
- Self-supervised and temporal self-attention with feature branching for real-time video polyp segmentation.

## 1.2 Thesis Structure and Contributions

This thesis presents an deep learning based framework with multi-tasks in colonoscopy. As shown in Fig. 1.3, the framework can monitor image quality and automatically detect lesions, and provide feedback to clinicians in real time. Assessment of image quality serves to standardise clinical procedures and offers an educational benefit to operators. Automatic detection of lesions can assist clinicians in diagnosis and effectively reduce the missed detection rate. The structure of the thesis is below:

- Chapter 1 introduces the clinical motivation of this thesis and the prominent role of deep learning-based image analysis methods for endoscopy. Then it summarises the challenges faced by clinicians in endoscopy and the corresponding solutions proposed in the thesis. This chapter also includes the thesis structure, contributions and related publications.
- Chapter 2 primarily delves into the clinical and technical background pertinent to this thesis. An understanding of the clinical context enriches the analysis of medical issues. Meanwhile, the technical background encompasses a comprehensive overview of classification, segmentation, and detection challenges within the realm of computer vision, with a particular focus on their application in endoscopic procedures.

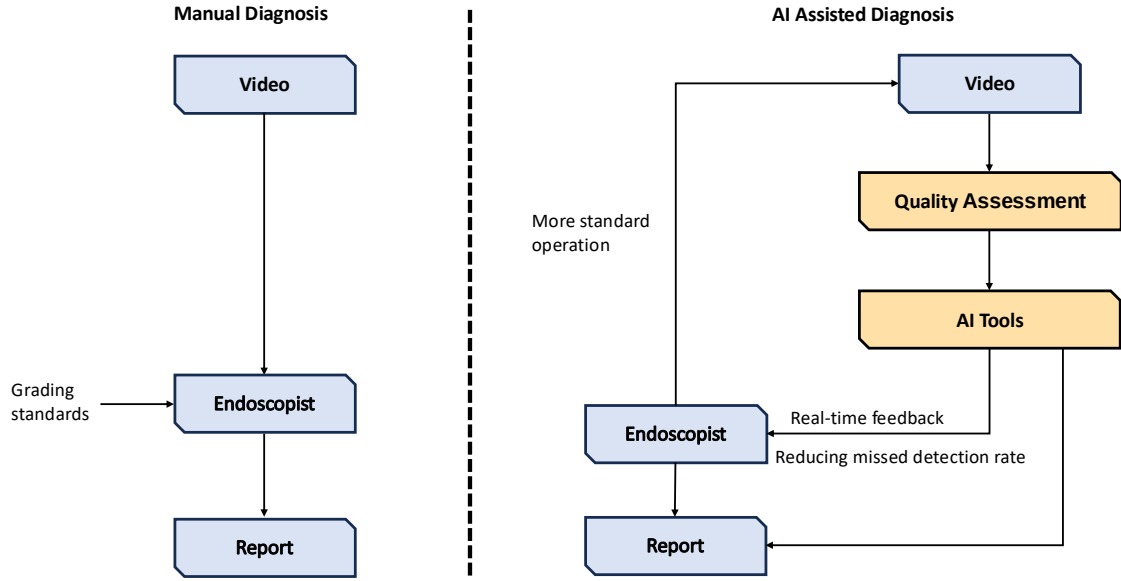


**Figure 1.3:** Overview of the thesis framework. The left column is the normal manual diagnosis. Proposed AI assisted diagnosis framework is on the right column.

- Chapter 3 presents an image quality assessment framework. Through our proposed novel bounding box pruning method, which can effectively improve artefact detection and provide high localisation scores of diverse artefact classes.
- Chapter 4 presents two deep learning-based UC grading methods. An end-to-end supervised method uses a model scaling network as the backbone and integration of additive angular margin loss. The proposed method can effectively increase the inter-class distance, thereby improving the accuracy of the UC grading task. In addition, a self-supervised learning framework called SSL-CPCD that explores patch-level instance-group discrimination and penalisation of inter-class variation using additive angular margin within the cosine similarity metrics is presented by improving generalisation ability and surpasses SOTA methods in multiple downstream tasks.
- Chapter 5 presents presents baseline methods, self-supervised approach proposed in Chapter 4 for polyp detection and segmentation task and a novel

end-to-end self-supervised and temporal self-attention framework for real-time video polyp segmentation.

- Chapter 6 encapsulates the principal contributions of this thesis and explores the limitations inherent to the current work, alongside potential challenges that may arise in future endeavors.

## 1.3 Publications

- **Xu, Ziang**, Sharib Ali, Soumya Gupta, Numan Celik, and Jens Rittscher. "Improved artifact detection in endoscopy imaging through profile pruning." In Medical Image Understanding and Analysis: 25th Annual Conference, MIUA 2021, Oxford, United Kingdom, July 12–14, 2021, Proceedings 25, pp. 87-97. Springer International Publishing, 2021.
- **Xu, Ziang**, Sharib Ali, James East, and Jens Rittscher. "Additive angular margin loss and model scaling network for optimised colitis scoring." In 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1-5. IEEE, 2022.
- **Xu, Ziang**, Sharib Ali, Soumya Gupta, Simon Leedham, James E. East, and Jens Rittscher. "Patch-level instance-group discrimination with pretext-invariant learning for colitis scoring." In International Workshop on Machine Learning in Medical Imaging, pp. 101-110. Cham: Springer Nature Switzerland, 2022.
- **Xu, Ziang**, Sharib Ali, Numan Celik, Adam Bailey, Barbara Braden, and Jens Rittscher. "Self-supervised Approach for a Fully Assistive Esophageal Surveillance: Quality, Anatomy and Neoplasia Guidance." In MICCAI Workshop on Cancer Prevention through Early Detection, pp. 14-23. Cham: Springer Nature Switzerland, 2022.

**Award:** Best paper award.

- **Xu, Ziang**, Jens Rittscher, and Sharib Ali. "SSL-CPCD: Self-supervised learning with composite pretext-class discrimination for improved generalisability in endoscopic image analysis." IEEE Transactions on Medical Imaging (2024).
- **Xu, Ziang**, Jens Rittscher, and Sharib Ali. "SSTFB: Leveraging self-supervised pretext learning and temporal self-attention with feature branching for real-time video polyp segmentation." arXiv preprint arXiv:2406.10200, 2024.(Neurocomputing-in review)

### **Co-authored publications**

- Gupta, Soumya, Sharib Ali, Ziang Xu, Binod Bhattarai, Ben Turney, and Jens Rittscher. "UNet-eVAE: Iterative Refinement Using VAE Embodied Learning for Endoscopic Image Segmentation." In International Workshop on Machine Learning in Medical Imaging, pp. 161-170. Cham: Springer Nature Switzerland, 2022.
- Pyatha, Ajay, Ziang Xu, and Sharib Ali. "Vision Transformer-Based Self-supervised Learning for Ulcerative Colitis Grading in Colonoscopy." In MIC-CAI Workshop on Data Engineering in Medical Imaging, pp. 102-110. Cham: Springer Nature Switzerland, 2023.

# 2

## Background

### Contents

---

|            |                                                 |           |
|------------|-------------------------------------------------|-----------|
| <b>2.1</b> | <b>Clinical background . . . . .</b>            | <b>11</b> |
| 2.1.1      | Gastrointestinal endoscopy . . . . .            | 11        |
| 2.1.2      | Gastrointestinal diseases monitoring . . . . .  | 12        |
| 2.1.2.1    | Upper gastrointestinal disease monitoring . . . | 12        |
| 2.1.2.2    | Lower gastrointestinal disease monitoring . . . | 13        |
| <b>2.2</b> | <b>Technical background . . . . .</b>           | <b>17</b> |
| 2.2.1      | Image classification methods . . . . .          | 17        |
| 2.2.1.1    | Traditional Methods . . . . .                   | 17        |
| 2.2.1.2    | Deep learning-based Methods . . . . .           | 18        |
| 2.2.1.3    | Medical image classification . . . . .          | 22        |
| 2.2.2      | Image detection methods . . . . .               | 25        |
| 2.2.2.1    | Traditional Methods . . . . .                   | 25        |
| 2.2.2.2    | Deep learning-based Methods . . . . .           | 26        |
| 2.2.2.3    | Artefact detection . . . . .                    | 29        |
| 2.2.2.4    | Polyp detection in endoscopy . . . . .          | 30        |
| 2.2.3      | Image segmentation methods . . . . .            | 31        |
| 2.2.3.1    | Traditional Methods . . . . .                   | 31        |
| 2.2.3.2    | Deep learning-based Methods . . . . .           | 32        |
| 2.2.3.3    | SOTA segmentation methods in colonoscopy .      | 34        |
| 2.2.4      | Relevant advances in deep learning . . . . .    | 36        |
| 2.2.4.1    | Self-supervised learning . . . . .              | 36        |
| 2.2.4.2    | Attention mechanism . . . . .                   | 41        |
| <b>2.3</b> | <b>Summary . . . . .</b>                        | <b>44</b> |

---

This chapter provides a comprehensive review of the literature relevant to this



thesis. This chapter is divided into clinical background and technical background. Initially, the clinical background section introduces gastrointestinal endoscopy and colorectal diseases involved in the thesis, including inflammatory bowel disease and polyps. Subsequently, the technical background section examines the three fundamental image analysis problems explored in the thesis: image classification, image detection, and image segmentation. Additionally, three specific research questions are discussed in different sections: 1) Medical image classification. 2) Artefact detection. 3) Polyp detection and segmentation.

## 2.1 Clinical background

This section presents a description of gastrointestinal endoscopy and gastrointestinal diseases which we focus on in the thesis.

### 2.1.1 Gastrointestinal endoscopy

Endoscopes are usually used by clinicians to examine hollow organs of patients [4]. The endoscope is composed of a camera with a light source and a correspondingly connected external monitor. The doctor can monitor the patient's body condition in real time through the monitor. Endoscopes can see lesions that cannot be shown by X-rays, so endoscopy has become an indispensable step. Endoscopes are mainly divided into two types: gastroscopy [5] and colonoscopy [6]. Gastroscopy is used to examine the esophagus, stomach and duodenum, while colonoscopy is used to observe the lesions of the rectum, colon and cecum. In addition, doctors can perform some treatments through endoscopes, such as the removal of benign polyps in the intestines. The earliest endoscopic exploration can be traced back to 1806, when a device invented by German doctor Philipp Bozzini used a candle as a light source and a series of mirror reflections to observe the inside of the human body. With the development of electronic technology, electronic endoscopes appeared in the 1980s [7]. This endoscope uses a miniature camera to provide higher resolution images, greatly improving the accuracy of diagnosis and the success rate of surgery. In recent years, wireless capsule endoscopes [8], using miniature cameras

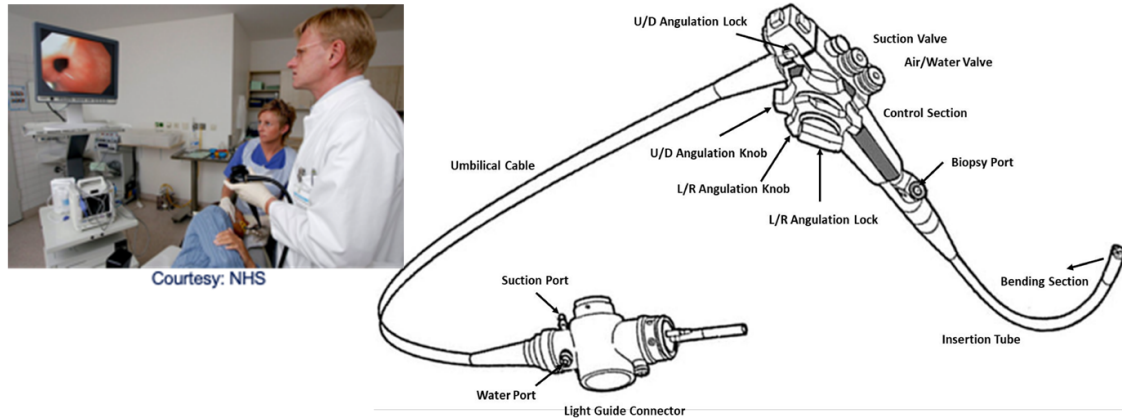


**Figure 2.1:** The endoscope is mainly composed of 4 major structural components, including control body, insertion tube, bending section/distal end and light guide connector [4]. Mechanical system, plumbing system and illumination system are respectively connected with various components, which provides a variety of functions for the endoscope.

and implanted in medical capsules, have also been used in clinical procedures. The wireless capsule endoscope can move in the direction of normal gastrointestinal muscle movement of the human body. Record data through the camera and transmit it to an external receiver. However, its main limitation is that it cannot be detected in real time and cannot be treated like a normal endoscope. Therefore, endoscopy is still the most widely used clinical procedure.

## 2.1.2 Gastrointestinal diseases monitoring

### 2.1.2.1 Upper gastrointestinal disease monitoring

Endoscopy is a routine clinical surveillance technique for diagnosis and minimally invasive treatments. Acid reflux disease, also termed as heartburn, is caused due to the acid flowing back into the stomach's food pipe. In gastroesophageal reflux disease, acid flows from the stomach into the food pipe, causing symptoms such as heartburn. Some patients with constant reflux over the years can develop a condition called Barrett's esophagus that is associated with a mildly increased risk of adenocarcinoma in the esophagus [9]. During upper endoscopy, a long, flexible tube with a camera at its tip is inserted through the mouth into the esophagus, and the lining of the gastrointestinal tract is observed on a separate monitor. This procedure lasts for about 20 min on average, during which the endoscopists visualise the esophagus,

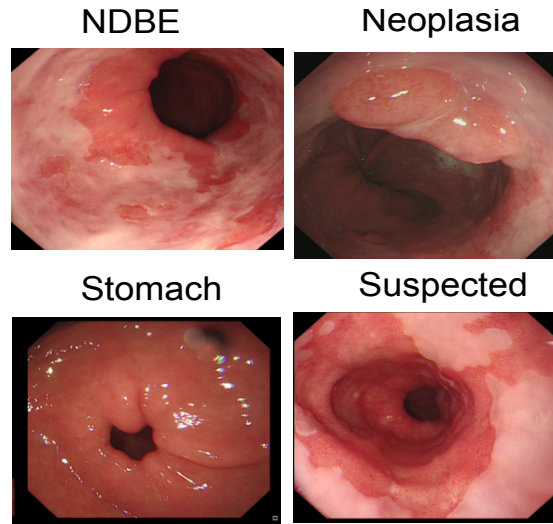


**Figure 2.2:** Samples of upper gastrointestinal diseases [11].

stomach, and upper part of the small bowel [10]. While doing so, the endoscopists try to understand and differentiate between the normal tissue (the normal squamous region in the esophagus), non-dysplastic Barrett's esophagus (NDBE), pre-cancerous neoplastic changes, and even esophageal adenocarcinoma (EAC).

#### 2.1.2.2 Lower gastrointestinal disease monitoring

The thesis focus on colorectal which widely used to stratify patients at risk to colorectal cancer. Two colorectal diseases are introduced below.

**Inflammatory bowel disease:** Inflammatory bowel disease (IBD) is a chronic, recurrent inflammation affecting the intestines. The most common IBDs include Crohn's disease (CD), ulcerative colitis (UC), etc. They affect the human body of any age, leading to long-term treatment and frequent surgical care [12]. Compared with Crohn's disease, ulcerative colitis mostly occurs in colon and rectum and has more focused features. The cause of ulcerative colitis is generally believed to be the result of a combination of genetics, environmental factors, abnormal immune system response, and an imbalance of intestinal microorganisms. Grading of UC is empirical, and the most commonly used grading systems in clinical practice are usually divided into two based on disease severity and disease extent. The Montreal Classification [13] divides UC into three categories by judging the extent of

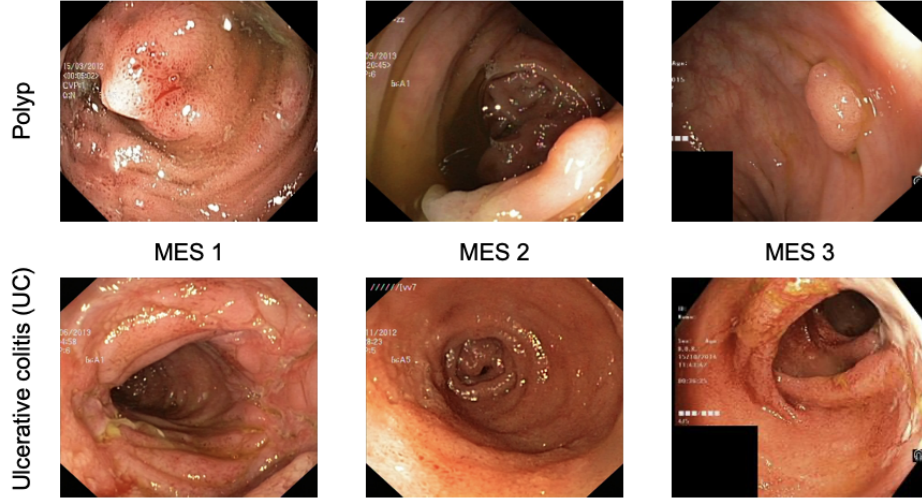


**Figure 2.3:** Images shows two classic diseases. The first line presents polyps and the second line presents 3 classes ulcerative colitis scored by Mayo endoscopic scoring.

inflammation covering the intestine, including (E1: anterior proctitis)inflammation limited to the rectum; (E2: left-sided colitis)inflammation extends to the left half of the transverse colon; (E3: generalised Colitis)Inflammation involving the entire colon. Mayo Clinic Endoscopic Subscore (MCES) [14] is a part of the Mayo scoring, which typically uses clinical symptoms to assess disease severity such as tool frequency, rectal bleeding, mucous membranes, etc. Mayo scoring defines UC as four categories: (Score 0)Normal or inactive disease; (Score 1)erythema, decreased vascular pattern, mild friability; (Score 2)marked erythema, absent vascular pattern, friability, erosions; (Score 3) spontaneous bleeding, ulceration. The presence of visible rectal bleeding and flexible sigmoidoscopic findings in Mayo scoring can be observed through clinical and endoscopy. At the same time, UC has relapse and remission problems.

Therefore, colonoscopy is the most suitable clinical tool for the examination and treatment of UC. Colonoscopy is used for the initial diagnosis of UC, and disease surveillance has been followed to assess treatment response. Colonoscopy is often performed to confirm the diagnosis of UC at the onset of symptoms (e.g., diarrhoea, rectal bleeding, and abdominal pain). During this initial procedure,

mucosal inflammation, ulcerations, and other visual characteristics are examined. In patients with a confirmed diagnosis of UC, colonoscopy is used regularly to assess disease activity and the degree of inflammation. This is particularly important when there is a sudden onset of symptoms and allows the clinician to assess whether the disease is confined to the rectum or involves a wider area of the colon. Colonoscopy is performed after initiating or adjusting therapy to determine the effectiveness of medications such as aminosalicylates and corticosteroids. In mucosal healing, the goal is to achieve visual improvement in inflammation corresponding to symptom control. However, endoscopic grading is often subjective. Despite the scoring systems, there can be variability between endoscopists in interpreting mucosal appearance, which may lead to inconsistent disease classification. Thus improving the reproducibility and objectivity of UC grading can greatly improve clinical decision-making. Artificial intelligence (AI)-powered tools can standardise the interpretation of endoscopic images, reducing inter-observer variability.

**Polyp:** Polyps are abnormal protrusions or small pieces of tissue that grow from the surface of tissue in the body. They vary in size, shape, and growth rate and may appear alone or in multiples [15]. The type of polyp depends on where they appear and the type of tissue they appear. Colon polyps are the most common, but they may also appear in other parts of the body, including the stomach, uterus, throat, cervix, and ear canal. Colon polyps that form on the mucosa of the colon or rectum are the main precursors of colorectal cancer. As the third most deadly cancer in the world, colorectal cancer is generally more common in people over the age of 50. Among them, having a history of polyps is considered an important risk factor for colorectal cancer. In addition, long-term lifestyle problems such as obesity, smoking, lack of physical activity, and inflammatory bowel diseases (such as Crohn’s disease, ulcerative colitis, etc.) also increase the risk of disease. Colon polyps can be divided into adenomatous polyps (with the potential to become cancerous) and non-adenomatous polyps (usually benign) according to their histological characteristics [16]. Colorectal cancer is

considered a relatively preventable malignancy. Colonoscopy can detect precancerous polyps in the early stages of cancer, and clinicians remove these early lesions during this process to prevent further development of cancer. Removing early cancer tissue through colonoscopic polyp removal surgery can avoid patients from undergoing subsequent chemotherapy, radiotherapy, and targeted therapy. Therefore, colonoscopy is considered the core of colorectal cancer treatment.

Colonoscopy is most often performed as part of routine colorectal cancer screening. This is usually recommended starting at age 45-50 or earlier, during the screening colonoscopy, the endoscopist searches for and removes adenomatous polyps, which are considered early detection and removal prevent the progression to colorectal cancer. After polyp removal, the endoscopist searches for and removes adenomatous polyps, which are considered precancerous. Although colonoscopy plays a key role in polyp detection, there are challenges associated with subjectivity and variability, similar to those in the grading of ulcerative colitis. Different endoscopists may have varying levels of skill in detecting subtle or flat polyps, leading to miss rates, especially for lesions. Different endoscopists may have varying levels of skill in detecting subtle or flat polyps, leading to miss rates, especially for lesions in challenging areas like the right colon. Adenoma detection rate (ADR) is an important quality metric for colonoscopies, but variability in ADR between endoscopists can result in under-detection, compromising patient outcomes. Higher ADR is associated with lower post-colonoscopy CRC rates. Therefore, to enhance clinical decision-making and improve patient outcomes, polyp detection in colonoscopy needs to become more objective and reproducible. To enhance clinical decision-making and improve patient outcomes, polyp detection in colonoscopy needs to become more objective and reproducible. AI-powered tools such as computer-aided detection (CADe), are being developed to assist endoscopists by highlighting suspicious areas in real time during the procedure. These technologies can increase ADR and standardise polyp detection, reducing the risk of missed lesions. These technologies can increase ADR and standardise polyp detection, reducing the risk of missed lesions.

## 2.2 Technical background

This section presents a detailed technical review on classification, detection and segmentation tasks in computer vision. And in particular, it gives a review of the research question of concern in the thesis. Finally, some recent advances in deep learning and computer vision have been provided to further trigger the novelty of this thesis.

### 2.2.1 Image classification methods

#### 2.2.1.1 Traditional Methods

Traditional image classification methods are divided into feature-based methods and template matching-based methods. Feature-based methods rely on manually extracted features, which are input into machine learning classifiers to classify images. Common features include:

- Color features: such as color histograms [17], color moments [18], etc., used to describe the color distribution in images.
- Texture features: such as gray-level co-occurrence matrix (GLCM) [19], local binary pattern (LBP) [20], etc., used to describe the texture information on the image surface.
- Shape features: such as contour descriptor [21], Hough transform [22], etc., used to describe the shape of objects in images.
- Spatial features: describe the spatial layout and positional relationship of objects in images.

A large number of extracted features may have high dimensions, so feature selection and dimensionality reduction are also very important. Commonly used methods include principal component analysis (PCA) [23] and linear discriminant analysis (LDA) [24], etc., which are used to reduce the dimension of features and select useful features. Finally, the selected features are input into the machine learning classifier to complete the image classification task. Support vector machine (SVM) [25] is

a supervised learning classification method that separates different categories by finding the optimal hyperplane in the feature space. SVM attempts to maximise the interval between different categories to achieve good generalization ability. Meanwhile, SVM is particularly suitable for classification tasks of high-dimensional data, and by using kernel techniques, it can effectively handle nonlinearly separable data. The K-nearest neighbor algorithm (K-NN) [26] is an instance-based learning method that does not require explicit model training. For a given input, the K-NN algorithm finds the K closest samples in the training data, and then predicts the category of the input sample based on the categories of these nearest neighbors. The decision tree [27] is a tree-structured classifier in which each internal node represents a test on an attribute, each branch represents the result of the test, and each leaf node represents a category. Classification rules can be obtained by following the path from the root node to the leaf node. Pruning techniques and ensemble methods (such as random forests [28]) are often used to improve the performance of decision trees. Random forests construct multiple decision trees and aggregate their prediction results to make the final classification decision. Random forests reduce the variance of the model and improve generalization ability by introducing randomness to construct each decision tree.

Although traditional image classification methods have achieved certain success, different feature extraction methods are required for different application scenarios, so expert knowledge is usually required to manually design and select features. In addition, for complex image scenes, traditional methods are easily affected by external factors such as lighting and noise. For high-dimensional image features, the method has a large computational overhead and is difficult to achieve efficient classification.

### 2.2.1.2 Deep learning-based Methods

With the development of deep learning technology in recent years, the Convolutional Neural Network (CNN) has become the dominant method for image classification tasks with its powerful feature extraction capabilities [31]. As the core of CNN, the convolution layer extracts features such as edges, corners, and textures from



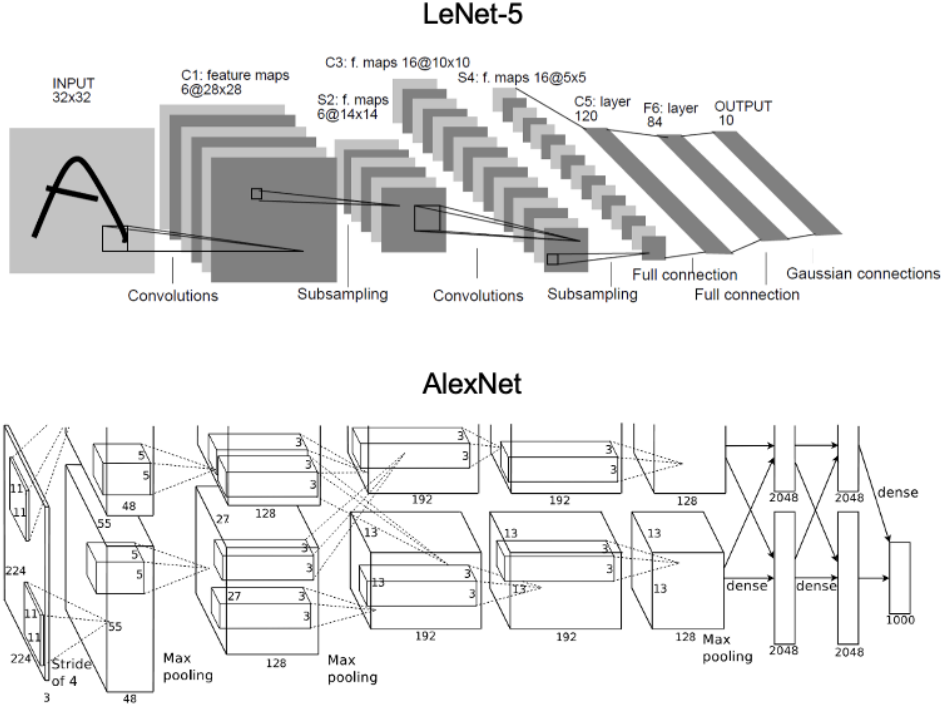


**Figure 2.4:** Architecture of LetNet-5 [29] and AlexNet [30]. The images intercept from the original paper.

the image by performing an inner product operation between local image regions at each spatial location and the convolution kernel, which is usually a  $3 \times 3$  or  $5 \times 5$  matrix. The extracted features are passed through a nonlinear activation layer, followed by a pooling layer that reduces the spatial resolution of the feature maps. The fully connected layer maps high-level features to the sample label space and outputs the probability of each category using the softmax function. LeNet-5 was a CNN architecture proposed by Yann LeCun et al. [29] in 1998. As shown in Fig. 2.4, it includes two convolutional layers, two pooling layers, and three fully connected layers and has been successfully applied to the task of classifying handwritten digits. Krizhevsky et al. [30] presented AlexNet for the classification task of large-scale image datasets. AlexNet consists of 5 convolutional layers, 3 pooling layers, and 3 fully connected layers. The rectified linear unit (ReLU) activation function helps the neural network converge quickly and alleviate gradient disappearance. At the same time, the Drop out technique is used as a method of model regularization to prevent overfitting. In the ImageNet Large Scale Visual Recognition Challenge

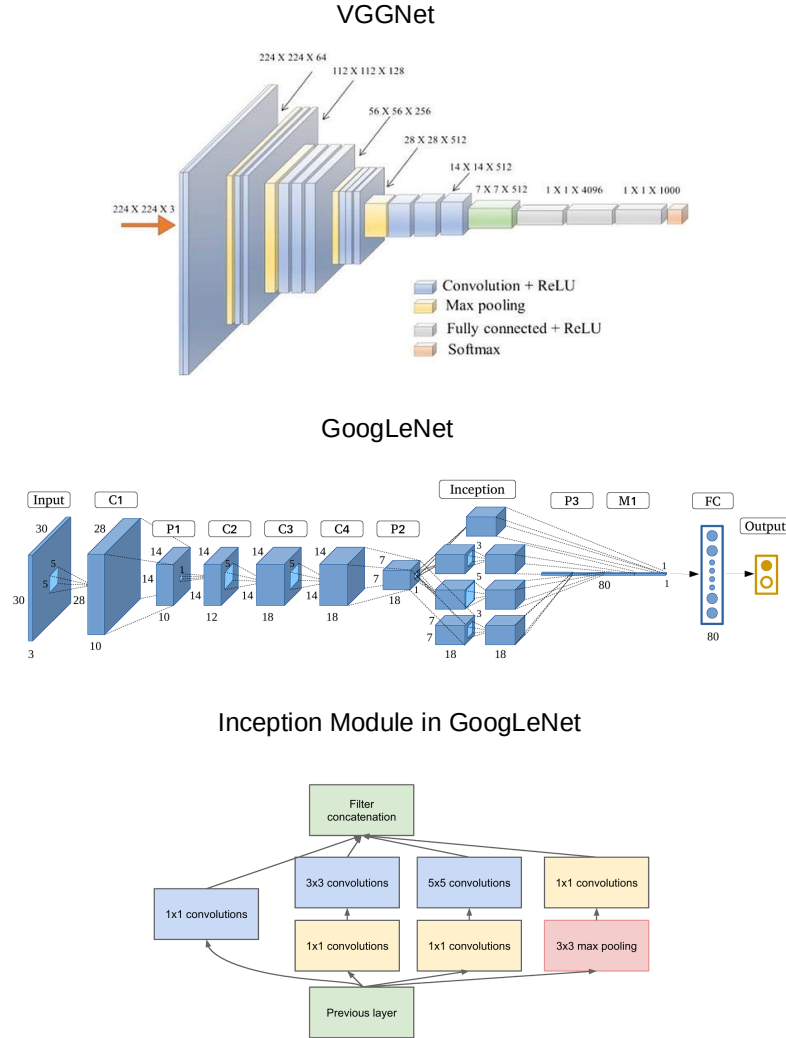


**Figure 2.5:** Architecture of VGGNet [33], GoogLeNet [34] and inception module. VGGNet and GoogLeNet utilise deeper network layers, in which Inception Module in GoogLeNet can not only maintain the sparsity of the network structure, but also perform high-performance calculations in dense matrices. The images intercept from the original paper.

(ILSVRC-2010) [32], AlexNet achieved 37.5% and 17.0% top 1 and top 5 error rates, exceeding the traditional method based on Fisher vector by about 8%.

In the following period of time, many studies focused on developing deeper network structures and larger model parameters in order to improve classification performance. Among the notable advancements, VGGNet proposed by Simonyan

and Zisserman [33], stood out by extending the network depth to 16-19 weighted layers and adopting multiple  $3 \times 3$  small convolutional kernels in place of the larger  $5 \times 5$  convolutional kernels used in AlexNet. In ILSVRC-2014, the top 1 and top 5 error rates were reduced to 23.7% and 7.3% respectively. However, VGGNet did not win the championship of ILSVRC-2014. GoogLeNet proposed by Szegedy et al. [34] reduced the top 5 error rate to 6.67%. GoogLeNet uses a larger and deeper network structure (22 layers), and proposes a module called Inception to find the optimal local sparse structure and combine different local structures together. The Inception module consists of  $1 \times 1$ ,  $3 \times 3$  and  $5 \times 5$  convolutional layers and  $3 \times 3$  maximum pooling layers, where the  $1 \times 1$  convolutional layer is used to calculate reductions before the  $3 \times 3$  and  $5 \times 5$  convolutions. GoogLeNet achieves compression and dimensionality reduction through the Inception module, reducing model parameters and thus saving computing costs. The architecture of VGGNet and GoogLeNet can be find in Fig. 2.5.

As the depth of the network continues to increase, researchers have found that gradient diffusion and explosion become obstacles to training deep networks, which makes the network unable to converge. The deep residual network (ResNet) proposed by Kaiming He et al. [35] solves this problem to a large extent. ResNet does not directly fit the target mapping, but indirectly learns the target mapping by fitting the residual mapping. Residual learning (as shown in Fig. 2.6) introduces Skip Connection, allowing the input to directly skip some layers and add the output of the skipped layer to form the final output. The introduction of residual learning allows the number of network layers to continue to increase. The 152-layer ResNet achieved a breakthrough in ILSVRC-2015, reducing the error rate of the top 5 to 3.57%. Based on ResNet, Gao Huang et al. [36] introduced a densely connected network (DenseNet), which uses dense blocks to connect all previous layers as additional input and splice input in the channel dimension. Unlike the hand-designed architecture of most models, the EfficientNet precented by Mingxing Tan and Quoc V. Le [37] utilising compound scaling to balance the depth, width and image resolution of the network at the same time. They are able to scale the model in

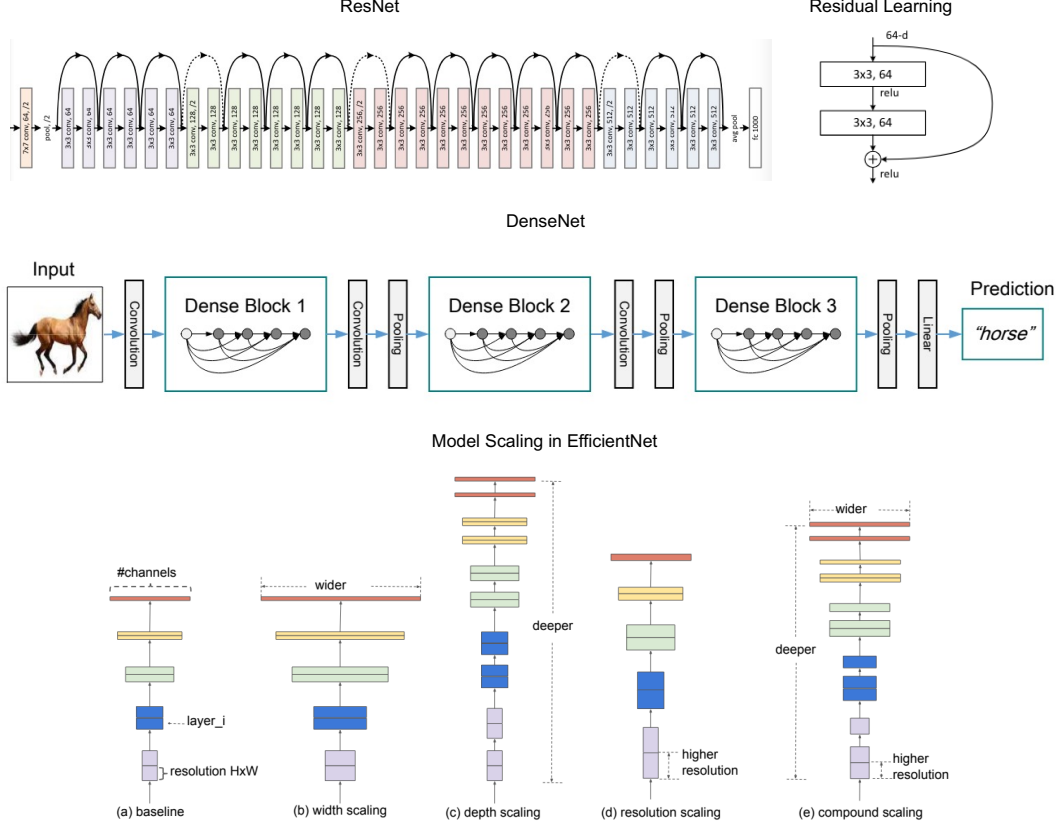


**Figure 2.6:** Architecture of ResNet [35], DenseNet [36] and model scaling in EfficientNet [37]. The residual learning in ResNet effectively solves the gradient vanishing problem in model training through identity mapping. DenseNet utilises the dense connection mechanism to allow the network to receive more information, while EfficientNet uses the Neural Architecture Search method to search for the best model structure. The images intercept from the original paper.

three dimensions at the same time in a balanced way, thereby effectively controlling the complexity and computational cost of the model while improving the accuracy.

### 2.2.1.3 Medical image classification

Medical image classification shares some of the same techniques and challenges as standard computer vision tasks, but differs significantly in terms of data constraints, accuracy requirements, and so on. Medical image datasets are typically much smaller compared to the large datasets used in standard computer vision, such as ImageNet. This is due to issues such as privacy concerns, labour and time costs involved in

annotating medical images. Meanwhile, there is a serious problem of category imbalance in medical datasets, which can lead to false negatives in the trained models and get serious consequences for diagnosis. In addition, medical images may have large variability due to factors such as imaging equipment, protocols, etc. This variability may reduce the ability of the model to generalise across institutions or patient groups.

Medical image Classification methods based on deep learning mainly use CNN. In [38], Zhou et al. established a GoogLeNet based on CNN to quantitatively measure the presence and extent of pathology in the entire small intestine. Park et al. [39] also used CNN to learn features on different scales. Ribeiro et al. [40] added data enhancement operations based on the use of CNN classification, such as translation, rotation, and scaling of the training data. 3-D Fully Convolutional Network (3D-FCN) was proposed by Yu et al. [41] for polyp classification in colonoscopy. They used a novel online learning program to solve the problem of limited training data by using specific information from the input video during the learning process. Due to the small amount of labelled data in the training set, over-fitting problems occur during deep learning model training. Data enhancement, parameter regularisation and transfer learning [42] are usually used to solve the problem of overfitting. In [43], Zhang et al. first used CaffeNet which to pre-train the natural dataset, and used transfer learning and fine-tuning techniques to adapt to the application of the medical dataset. Transfer learning is also used in a variety of medical image classification tasks [44, 45].

### **Disease classification in gastrointestinal tract**

Upper gastrointestinal tract: Various studies have been explored in the literature that aimed at the classification and detection of NDBE, dysplastic BE, and EAC using deep learning methods [46–48]. A deep learning-based computer-aided method using ResNet18 was used to classify early neoplasia and NDBE [46]. The classifier showed an accuracy of 89% accuracy with 90% sensitivity and 88% specificity on the test set. Furthermore, the model outperformed the general endoscopists for a

classification task by 10-20% in accuracy, sensitivity, and specificity. Liu et al. [47] reported a CNN-based model for the classification of esophageal cancer (EAC) from early neoplasia and normal endoscopic frames. The proposed approach was based on the Inception-ResNet model and comprised of two sub-networks: O-Net for extracting the color and global features and P-Net for extracting the texture and detail features. Their approach achieved an accuracy of 85.83%, a sensitivity of 94.23%, and specificity of 94.67% after the fusion of the two subnetworks. Similarly, a deep learning approach for classification EAC (often obvious lesion) using ResNet and NDBE was developed by Ebigbo et al. [48]. WLI and NBI modalities were used, and the authors reported a sensitivity of 97% and specificity of 88%. A dual-stream network for classification and segmentation of esophageal lesions was proposed by Wu et al. [49]. A large image dataset of 189,436 images was used by Hussein et al. [50] for dysplasia (high-grade dysplasia or EAC) and NDBE classification using ResNet-101. For the anatomical site classification task, He et al. [51] used 3704 images of upper GI endoscopy for training and evaluated the classical models. Their best model using DenseNet-121 provided an overall accuracy of 88%. A similar work to [51] by Park et al. [52] employs several classical CNN-based models to classify anatomical sites that included the esophagus, stomach, and duodenum; for which 11 different annotated sub-classes were present. Their best-performing model was EfficientNet-B1. GoogLeNet was used by Takiyama et al. [53] to classify anatomical sites with a sensitivity of 96.4% and specificity of 98.1%.

Lower gastrointestinal tract: Most of the research work on UC classification is based on CNN. The Mayo Endoscopy Score is one of the most widely used and validated indicators to determine the treatment of UC. In these studies (Ozawa et al. [54], Stidham et al. [55], Alammari et al. [56]), they evaluated the feasibility of CNN to classify the severity of UC in still frames. Gutierrez et al. [57] proposed a grading system for the severity of UC, which also uses the Mayo endoscopy score as the scoring standard. The method includes a quality control model whose purpose is to distinguish between readable and unreadable frames, and a deep learning network based on CNN for UC classification. Ozawa et al. [54]. also utilised CNN

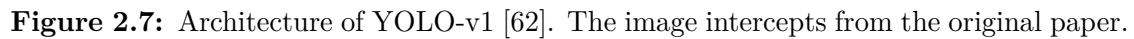
as a classification model, and divide them into three categories according to the severity of UC, namely Mayo 0, Mayo 1 and Mayo 2-3.

## 2.2.2 Image detection methods

### 2.2.2.1 Traditional Methods

Similar to classification methods, traditional image detection algorithms are also based on manual feature extraction and machine learning classifiers. Here are some classic methods applied to object detection:

- The VJ detector [58] adopts the most direct detection method, using a sliding window to check whether all possible positions and scales in the image contain the detection target. The integral image method is used to accelerate the filtering or convolution calculation, making the computational complexity of each window in the VJ detector independent of its window size. Then the Adaboost algorithm is used to select the most useful features from a group of random feature pools. The cascade classifier reduces the amount of calculation of the background window through the strong and weak classifier structure, thereby optimizing the computational efficiency.
- The HOG+SVM detector [59] employs Histogram of Directed Gradients (HOG) feature [60] combined with a linear SVM classifier to detect pedestrians and other objects in a sliding-window approach. HOG [60] is a widely used feature descriptor for object detection. It captures local shape and gradient information by computing gradient orientation histograms over dense, uniformly spaced cells.
- The Deformable Part-based Model (DPM) [61] extends the HOG-based approach by modelling objects as collections of parts arranged in a deformable configuration. Each part is described by a HOG template, and the spatial relationships between parts are learned using latent SVMs in a weakly supervised setting. DPMs enable improved robustness to object variation and deformation.



CNN has brought the field of object detection to a new level with its powerful feature extraction capabilities. CNN-based object detection algorithms are mainly divided into anchor-based and anchor-free methods. Anchor-based methods rely on a predefined set of anchors with different sizes and aspect ratios for scanning at different locations in the image to find possible targets. The method first places multiple anchors at different locations in the image, and then classifies and regresses each anchor to determine the category and precise location of the target object. Unlike anchor-based methods, anchor-free methods do not rely on predefined anchors to locate objects. Instead, anchor-free methods directly predict the existence, size, shape, or other attributes of objects in the image, or predict the key points of the target (such as the center point or corner point of the target object).

A holistic deep learning framework for endoscopy



the grid cell is responsible for predicting the target. YOLO-v1 presented 52.7% of mAP and achieved a real-time performance of 155 fps on the VOC-07 dataset [64]. The difference between the SSD algorithm and YOLO-v1 is that YOLO-v1 only performs detection in the deepest branch of the network, while SSD has multiple different detection branches, and different detection branches can detect objects of multiple scales. Therefore, SSD has greatly improved the accuracy of multi-scale object detection, increasing mAP to 76.8% on the VOC-07 dataset. RetinaNet introduced by Lin et al. [65] is another single-stage detector where the authors introduced focal loss for automatically adjusting weights to replace the standard cross entropy loss making the model training more focused on difficult samples. RetinaNet consists of a backbone (Resnet), a feature pyramid net (FPN) and two FPN subnets (class subnet and box subnet). FPN enhances the multi-scale features extracted from resnet to obtain a more expressive feature map containing multi-scale target area information. Finally, two FPN subnets with the same structure but no shared parameters are used on the feature map of FPN to complete the target box category classification and box position regression tasks. EfficientDet [66] uses EfficientNet [37] as the feature extraction network and introduces the Bidirectional Feature Pyramid Network (BiFPN) to fuse multi-scale feature maps from top to bottom and bottom to top, thereby achieving higher-level feature fusion.

On the other hand, multi-stage detectors first generate the region proposal, and then perform object detection for the selected object known as *region proposals*. Among them, two-stage architecture such as R-CNN [67], Fast R-CNN [68] and Faster R-CNN [69] have become the basis of many successful models in the field of object detection [70–72]. R-CNN uses the Selective Search algorithm to extract about 2,000 Region Proposals that may contain objects from bottom to top in the image. Each Region Proposal is warped into a uniform size and input into CNN, and the output of the fully connected layer is input into SVM for classification. Compared to R-CNN, Fast R-CNN adds an ROI pooling layer after the last convolutional layer to map inputs of different sizes to a feature vector of a fixed scale. In addition, softmax is used instead of SVM, and the bounding box regression task is added

to the training process using a multi-task loss function, making the entire training process end-to-end. Faster R-CNN proposes a Region Proposal Network (RPN) to replace Selective Search, which provides more accurate bounding box location information through a sliding window. Cai and Vasconcelos proposed Cascade R-CNN [73], which further stacked several cascade modules in the detector part and used different IOU threshold training to improve the positioning accuracy of the detector. Cascade R-CNN has brought the accuracy of the two-stage object detection algorithm to a new level, and achieved 42.8% mAP in the COCO dataset [74], exceeding other SOTA methods in the same period.

In anchor-based methods, the size, number, and aspect ratio of anchors have a huge impact on detection performance. For different tasks, anchors must be constantly resized, which greatly affects the generalization ability of the model. Therefore, researchers have begun to focus on anchor-free methods that do not rely on anchors. CornerNet [75] proposes a new object detection method that converts the network’s detection of the target bounding box into the detection of a pair of key points (i.e., the upper left corner and the lower right corner). The method uses a single convolutional network to predict a heat map and an embedding vector for each detected corner point for the upper left corner and the lower right corner of all instances of the same target category. By learning the vector embedding between corner points, the corner points belonging to the same object have similar embedding vectors. CenterNet [76] abandons the idea of predicting the upper left corner and the lower right corner key points, and directly detects the center point of the target. The heat map of the center point of each object is predicted by the convolutional neural network, and the center point of each object corresponds to a peak on the heat map. At the same time, the width and height of the object and the precise offset of each center point relative to its grid cell are predicted, which is used to determine the bounding box position of the object. On the COCO dataset, CenterNet achieved a mAP of 47.0%.

### 2.2.2.3 Artefact detection

The artefact detection task is a key step in image quality assessment, as well as the pre-process for the entire endoscopic image analysis framework. Therefore accurate artefact detection is crucial for the further task. Researchers typically use single-stage detectors or multi-stage detectors for artefact detection.

Single-stage detectors, such as YOLO-v1 [62] and SSD [63], refer to detectors that directly regress the category probability and position coordinates, and only perform a single pass on the data. RetinaNet [65] is another single-stage detector where the authors introduced focal loss that puts a focus on hard examples to boost performance. Some researchers adopt single-stage detectors for artefact detection tasks. Subramanian et al. [77] used RetinaNet model with ResNet50 and ResNet-101 feature extractors for the detection of artefacts. Oksuz et al. [78] and Khan et al. [79] utilised the RetinaNet model and focal loss to detect artefacts in artefact detection challenge EAD2019 [70].

However, single-stage detector based methods are usually less accurate than multi-stage methods. Multi-stage detectors first generate the region proposal, and then perform object detection for the selected object known as *region proposals*. Therefore, more researchers utilise multi-stage architecture such as R-CNN [67] and Faster R-CNN [69] which become the basis of many successful models in the field of object detection on medical image [70–72]. Polat et al. [71] combined Faster R-CNN, Cascade R-CNN and RetinaNet. The prediction of their model is based on class agnostic non-maximum suppression and false-positive elimination. Yang et al. [80] improved Cascade R-CNN model by combining feature pyramid networks (FPN). By fusing multi-scale feature representation and multi-stage object detection through cascade structures, their method can detect artefacts efficiently. Zhang et al. [72] implemented a mask-aided R-CNN model with a flexible and multi-stage training protocol to deal with artefact detection tasks. To overcome the problem that models trained at the same scale do not work well when detecting objects of the same category at different scales, Wang et al. [81] proposed a data preprocessing method to cut patches containing the same object from the image for training by

setting a threshold. For the detection task, they used a Faster R-CNN with a FPN as the backbone. With Mask R-CNN, Watanabe et al. [82] leverage available image segmentation annotations to help with bounding box detection, thereby improving the performance of YOLOv3. Yuan et al. [83] presented a multi-scale detection framework based on YOLOv5 called EADYOLO. The multi-scale detection layer and the dilated space pyramid pooling module are used to enhance the feature extraction capability of the encoder. Depthwise separable convolution can reduce the number of parameters while maintaining accuracy.

#### 2.2.2.4 Polyp detection in endoscopy

Deep learning-based object detection algorithms are also widely used in polyp detection task. Lee et al. [84] used YOLOv2 and validated the algorithm on public datasets and colonoscopy videos, demonstrating real-time capability as one of the milestones. Zhang et al. [85] proposed a Single Shot MultiBox Detector (SSD) for gastric polyps. They linked the feature maps from the lower layers, and the feature maps deconvolved from the upper layers and improved the mean precision (mAP) from 88.5% to 90.4%. Qadir et al. [86], and Shin et al. [87] used Mask R-CNN and Faster RCNN with different backbones to detect polyps, respectively. To address the variations among different datasets, Liu et al. [88] introduced a Gaussian Fourier domain adaptation method. This approach involves sampling matching pairs of source and target images from a Gaussian distribution and then standardizing their style through low-level spectral replacement. Such a process can minimise the domain differences in appearance across cross-device polyp datasets without altering their content. The hierarchical attention adaptation module reduces the domain differences in feature semantics, enabling the method to achieve a recall of 87.5% in the extended polyp detection task. The ENDOMIND-Advanced system was proposed by Krenzer et al. [89] for real-time automatic polyp detection. They organised and merged open source polyp detection datasets to create a dataset of more than 500,000 annotated images. The method used

YOLOv5 as the detection architecture and achieved an F1 score of 90.24% in the CVC-VideoClinicDB benchmark.

### 2.2.3 Image segmentation methods

#### 2.2.3.1 Traditional Methods

Traditional image segmentation divides regions by grayscale, color, texture, and shape features in the image, maximizing the differences between regions and minimizing the similarities within regions. It mainly includes the following:

- Threshold-based segmentation method [90]: One or more grayscale thresholds are calculated based on the grayscale features of the image. After the grayscale value of each pixel in the image is compared with the threshold, the pixel is divided into the appropriate category according to the comparison result. Therefore, the most critical part of the threshold method is to solve the optimal grayscale threshold according to the criterion function. Commonly used criterion functions include fixed threshold method, histogram bimodal method, iterative threshold method and adaptive threshold method, etc.
- Edge-based segmentation method [91]: The discontinuity of local image features or the sudden change of grayscale, color, texture and other characteristics can usually be observed on the boundary lines of different regions. The edge-based segmentation method refers to edge detection based on grayscale values, which is based on the observation that the grayscale value of the edge will show a step-type or roof-type change. There is a significant difference in the grayscale values of the pixels on both sides of the step-type edge, while the roof-type edge is located at the turning point where the grayscale value rises or falls. Based on this characteristic, differential operators can be used for edge detection, that is, the extreme value of the first-order derivative and the zero point of the second-order derivative are used to determine the edge. Common detection algorithms include Canny edge detector, Harris corner detector and SIFT detector, etc.

- Region-based segmentation method [92]: Divide the image into different regional blocks according to the similarity measure. The seed region growing method starts with a set of seed pixels representing different regions, and gradually merges similar pixels around the seeds to expand the region until it cannot be merged. The similarity within the region can be measured using information such as average gray value, texture, and color. The region splitting and merging method divides the image into multiple non-intersecting regions, and splits and merges these regions according to pre-established criteria. The watershed method is a mathematical morphological segmentation method based on topological theory. The gray value of each pixel in the image represents the altitude of the point. By simulating the process of water level rise, when water in different regions meets, a watershed is established between them. Repeat this process until the entire image is completely submerged, and the series of watersheds established at this time become the boundary lines separating the image regions.

In addition to region growing, other classical region-based approaches include contour-based models such as Active Contour Models (Snakes) [93] and Level Set Methods [94]. These techniques represent region boundaries explicitly or implicitly and evolve them over time by minimising energy functions based on image gradients and region homogeneity. They are particularly useful for accurately segmenting objects with smooth or complex boundaries, and can naturally handle topological changes like splitting or merging.

### 2.2.3.2 Deep learning-based Methods

In recent years, image segmentation methods based on deep learning have received considerable attention from researchers. The Fully Convolutional Network (FCN) was proposed by Long et al. [95] in 2015 and has become the milestone for the pixel-level segmentation task. As illustrated in the Fig. 2.8, unlike CNN, which uses a fully connected layer to obtain a fixed-length feature vector for classification and a deconvolution layer to upsample the feature map of the last convolution

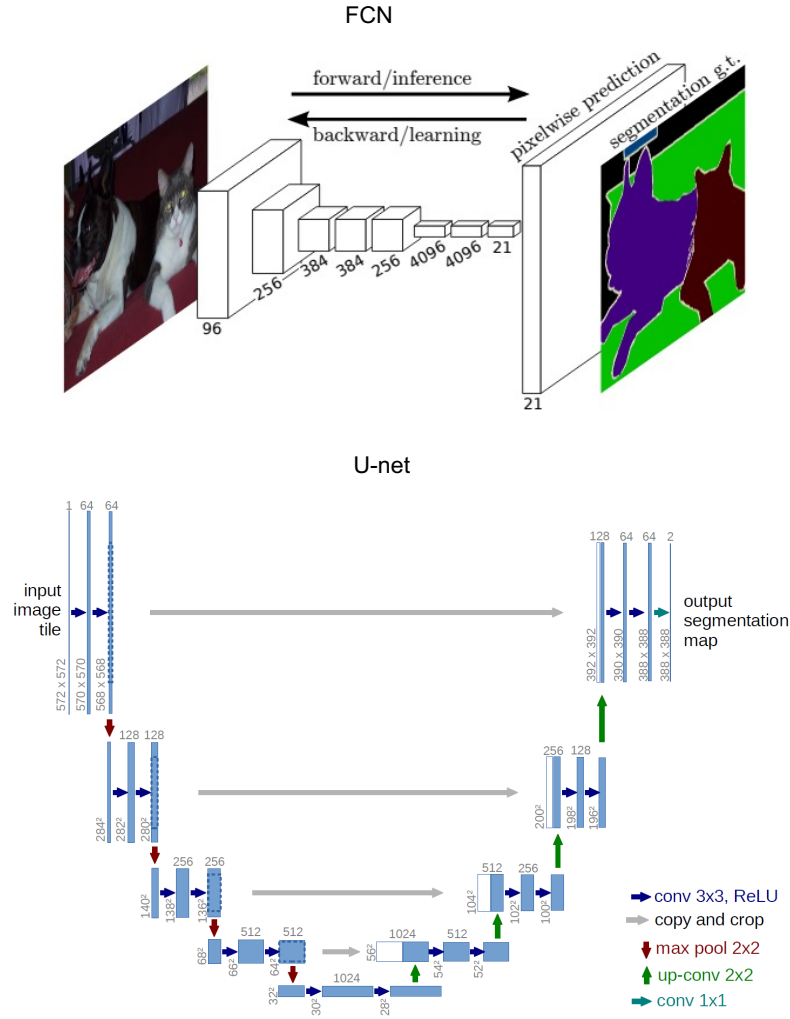


**Figure 2.8:** Architecture of FCN [95] and U-net [96]. The images intercept from the original paper.

layer to restore it to the same spatial size as the input image. Therefore, FCN can generate a prediction for each pixel while retaining the spatial information in the original input image. Similar to FCN, U-net [96] also utilises the upsampling and downsampling structure, which is also called an encoder-decoder architecture. The left side of the architecture (Contracting Path) contains multiple convolutional layers and maximum pooling layers to reduce the size of the feature map and double the number of feature channels. The right side of the architecture (Expansive Path) uses deconvolution layers to upsample the feature map to gradually increase the size of the feature map to the same as the input image. A key feature of U-Net is the

skip connection, which splices the corresponding feature maps in the contracting path with those in the expanding path. This enables the model to retain more spatial information, helping it to be more accurate when performing pixel-level segmentation. U-net is widely used in biomedical image segmentation tasks as a baseline due to its performance. Extensions of the original U-Net [96] such as U-Net++ [97], ResUNet [98] and Dense U-Net [99] were applied to the biomedical image segmentation and demonstrated improved accuracy.

### 2.2.3.3 SOTA segmentation methods in colonoscopy

**Frame based polyp segmentation:** The visual similarity of polyps and the colonic mucosa severely limits the utility of hand-crafted features utilised in early approaches [100, 101]. Deep learning-based approaches for polyp segmentation [41, 102] have helped to overcome these limitations. PsiNet [103] focused on region boundary constraints, exploiting polyp region and boundary information. In addition, Fang *et al.* [104] proposed an aggregated multi-scale and multi-sensory field feature aggregation network for polyp segmentation and considered the dependence between region and boundary branches through boundary-sensitive loss functions. PraNet [105] achieves cutting-edge segmentation accuracy and generalisation performance by mining the relationship between regions and boundary cues through parallel reverse attention modules. The method improved with mean Dice similarity coefficient (mDice) over 8% and 6% on the Kvasir and CVC-ClinicDB datasets, respectively. Polyp-PVT [106] adopted the transformer encoder, which can learn more robust representations. The three modules incorporated in the method are employed for gathering semantic and positional information of polyps from high-level features, capturing polyp information features in the lower layers, and extending pixel features of polyp regions with semantic positional information, facilitating a more effective feature fusion. Similarly, the recently proposed PVT-CASCADE [107] and CTNet [108] also utilised the transformer structure. In order to enhance the model’s ability to learn the relationship between contexts among pixels, PVT-CASCADE proposed the CASCADED Attention module, which enhances



long-distance and local dependencies by suppressing the background and fuses features with skip connections. Multi-scale interaction modules are used in CTNet to acquire high-resolution feature maps and multi-scale information. Ta *et al.* [109] devised a boundary learning and enhancement network that integrates edge details into higher-level features through a fusion approach, thereby generating features for segmentation with boundary information. Subsequently, multi-level boundary-aware attention blocks provided ambiguous edge information and enhanced the previous segmentation results. Qiu *et al.* [110] proposed an encoder-decoder network called Boundary Distribution Guided Network (BDG-Net). In the encoder, Gaussian distribution is employed to extract further information about the boundaries of polyps, which is then fused with low-level and high-level features. However, due to its heavy reliance on boundary features, this method exhibits sub-optimal segmentation results when polyps have weak edges. [111] also focuses on polyp boundary information. They proposed a multi-scale global perception approach for aggregating multi-scale information and expanding the receptive field to acquire the primary distribution of polyps at local and global levels. Experimental results in the paper indicated that their proposed method outperforms most state-of-the-art models on five publicly polyp datasets. CoInNet [112] proposed by Jain *et al.* focused more on tiny polyps. They designed a novel feature extraction strategy that utilises statistical information between pixel intensities on the feature map. This information was then input into convolutional layers to highlight the polyp regions in the image. The approach achieved mDice of 0.926 and 0.930 on Kvasir and CVC-ClinicDB datasets, which reached state-of-the-art levels. Zhou *et al.* [113] presented a cross-level feature aggregation network called CFANet. They performed cross-layer feature fusion at different scales by fusing features of adjacent layers, and merged boundary information.

**Video based polyp segmentation:** Zhao *et al.* [114] adopted a temporal local context attention mechanism to refine the prediction of the current frame through the prediction results of nearby frames, and a neighbouring frame spatiotemporal

attention mechanism used to capture long-range dependencies in time and space. The experimental results in the paper showed that the semi-supervised method they proposed achieves similar results to the fully supervised method. Self-attentive networks [115] can capture the internal relevance of features and reduce the dependence on external information. Ji *et al.* [116] proposed PNS-Net using self-attention blocks, equipping with recurrence and CNNs to improve polyp segmentation accuracy and achieve real-time inference speed. They further extended their approach [117] focused on video polyp segmentation, which accomplished maximum Dice index of 0.756 and 0.737 in SUN-SEG-Easy and SUN-SEG-Hard. The recently proposed ESFPNet [118] utilised a pre-trained Mix Transformer (MiT) encoder and an efficient staged feature pyramid (ESFP) decoder structure to achieve accurate image segmentation. Although the method focused on bronchial lesion segmentation, it achieved excellent results on the polyp segmentation dataset. The approach obtained mDice of 0.811 and 0.823 in CVC-ColonDB and ETIS-LaribPolypDB, respectively. Xiao *et al.* [119] proposed ICBNet based on iterative feedback learning strategy by imitating the working paradigm of clinicians. ICBNet gradually improves boundary detail information and context segmentation results in an iterative manner. Experimental results show that the proposed method can obtain accurate and robust results.

## 2.2.4 Relevant advances in deep learning

### 2.2.4.1 Self-supervised learning

Self-supervised learning (SSL) uses pretext tasks to mine self-supervised information from large-scale unsupervised data, thereby learning valuable image representations for downstream tasks. By doing so, the limitation of network performance on predefined annotations is greatly reduced. Learning based on pretext tasks helps to overcome the limitations of supervised learning by making more efficient use of the available data and reducing the reliance on labelled datasets. SSL is usually divided into two types: contrastive learning based methods (Fig. 2.9) and generative based method (Fig. 2.10). In contrastive learning methods, the pretext task usually

applies a transformation to the input image and predicts the transformed properties from the transformed image. During pretext learning, a single model[120–122] is used for both image and its transformations to learn consistent, invariant, and semantically meaningful representations. The same model is then used for fine-tuning on downstream tasks. The essence of contrastive learning is to learn a representation learning model by constructing similar and dissimilar instances, so that similar instances are closer in the projection space, while dissimilar instances are farther away in the projection space. The concept of contrastive loss was initially introduced by Bromley et al.[123]. Subsequently, more formal definitions were proposed by Chopra et al.[124] and Hadsell et al. [125], further refining the concept. For a pair of images  $(x_i, x_j)$ , their labels are  $y_i$  and  $y_j$  respectively.  $f_\phi()$  extracts the feature embedding of samples. Where  $m$  is a hyperparameter that defines the lower bound of the distance between different samples. The contrastive loss is defined as:

$$\begin{aligned} \mathcal{L}_{\text{contrastive}}(x_i, x_j, \phi) = & [y_i = y_j] \|f_\phi(x_i) - f_\phi(x_j)\|^2 \\ & + [y_i \neq y_j] \max(0, m - \|f_\phi(x_i) - f_\phi(x_j)\|)^2 \end{aligned} \quad (2.1)$$

If samples are similar( $y_i = y_j$ ), then minimise the Euclidean distance between them. If samples are not similar, then maximise the Euclidean distance between them up to a limit  $m$ .

Based on contrastive loss, researchers have proposed a large number of self-supervised learning framework in recent years. Chen et al. [121] proposed the SimCLR model, which is a typical paradigm for contrastive learning of visual representations (as demonstrated in Fig. 2.9). SimCLR performs data enhancement on the input image to simulate the input from different perspectives of the image. A contrastive loss is then used to maximise the similarity of the same object under different data augmentations and minimise the similarity between similar objects. Later, the MoCo model proposed by He et al. [122] also used contrastive loss to compare the similarity between a query and the keys of a queue to learn feature representation. The authors used a dynamic memory, rather than a static memory bank, to store feature vectors used in training. In contrast to these

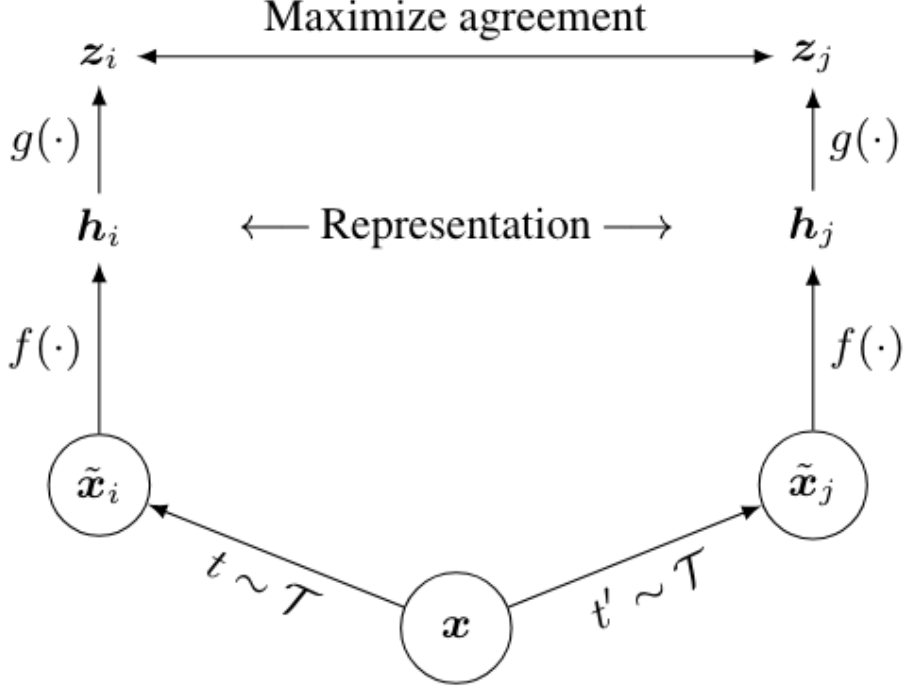


**Figure 2.9:** Typical paradigm in contrastive learning of visual representations [121]. The original input image  $x$  undergoes two different data augmentations ( $t \sim T$  and  $t' \sim T$ ) in the same series to generate two related views ( $\tilde{x}_i$  and  $\tilde{x}_j$ ).  $f(\cdot)$  and  $g(\cdot)$  are the encoder network and projection head, respectively. The agreement between feature embeddings ( $z_i$  and  $z_j$ ) is increased by minimizing the contrastive loss during the training phase. The representation  $h$  is used for downstream tasks.

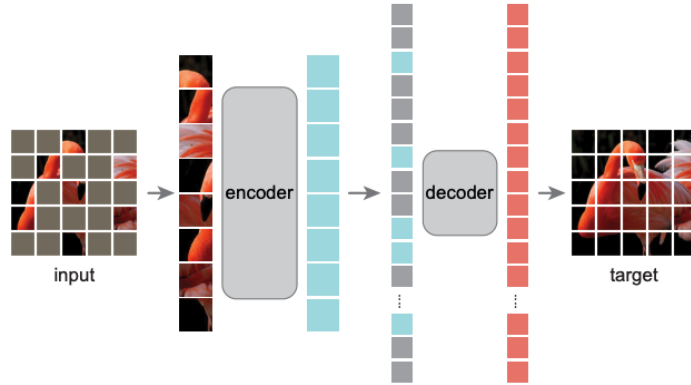


**Figure 2.10:** Architecture of generative self-supervised learning [126].

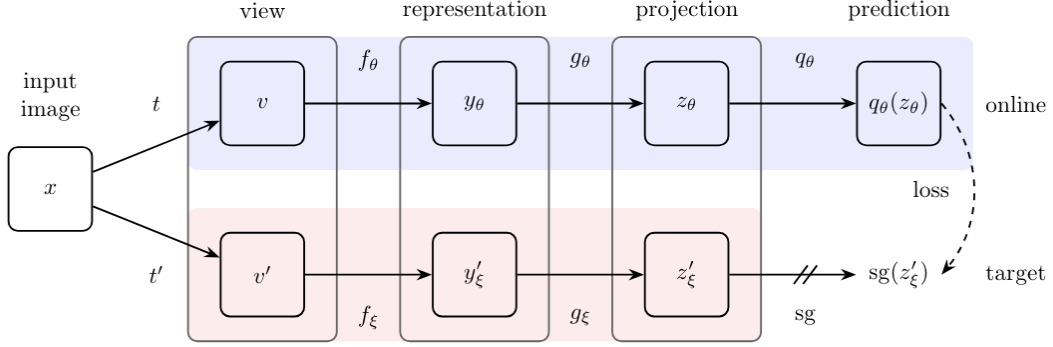


**Figure 2.11:** Architecture of BYOL [129]. Online network and target network have similar structure with encoder  $f$  and projector  $g$ . The difference is that the online network updates parameters  $\theta$  through training, and the parameters  $\xi$  of the target network are an exponential moving average of the online parameters. The mean squared error is minimised between the online network output through the predictor  $q_\theta(z_\theta)$  and the target network output through the stop-gradient  $sg(z'_\xi)$ . The image intercepts from the original paper.

methods that encourage the construction of covariant image representations to the transformations, pretext-invariant representation learning (PIRL) [120] pushes the representations to be invariant under image transformations. PIRL computes high similarity to the image representations similar to the representation of the transformed versions and low resemblance to representations for the different images. The notion of a Jigsaw puzzle [127] was used as a image patch representation for PIRL representation learning. Wang et al. [128] integrate between-instance similarity into contrastive learning, not directly by instance grouping, but by cross-level discrimination (CLD) between instances and local instance groups. Incorporated between-instance similarity into contrastive learning through cross-level discrimination (CLD) without relying on direct instance grouping. The CLD approach involved discerning similarities between instances and local instance groups. The proposed CLD can significantly improve the positive or negative sample ratio of contrastive learning, achieve better invariant mapping, and be embedded as an add-on component in other self-supervised methods.

Unlike the methods proposed above, BYOL (Bootstrap Your Own Latent) [129] does not require negative samples for contrastive learning. Instead, they learn useful features by maximizing the consistency between the representations of two different

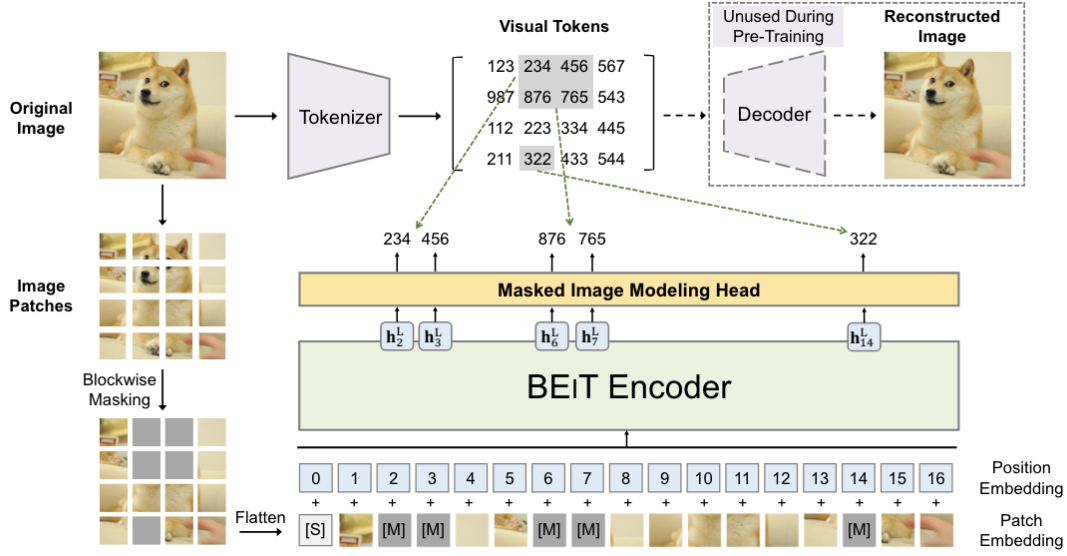


**Figure 2.12:** Architecture of BEiT [132]. The strategy for pre-training involves reconstructing the original visual tokens from the corrupted image patches. The image intercepts from the original paper.

views (two variants of the same image generated by data augmentation). As shown in Fig. 2.11, BYOL’s network architecture consists of an online network and a target network. The feature representation of the online network is converted into a predicted representation by the predictor, and then the Euclidean distance between the online network predicted representation and the target network representation is minimised. The online network is trained and updated by backpropagation, while the weights of the target network are updated by the exponential moving average of the weights of the online network. DINO (Self-Distillation with No Labels) [130] also uses self-distillation and replaces the encoder with a vision transformer [131]. A bias is added to the target network for centralization to avoid excessive sensitivity to small batch data training.

In generative based method, researchers usually adopt the masked image modeling (MIM) strategy, which enables the model to learn the effective features of the image by predicting the masked part of the original image [126]. BEiT was proposed by Bao et al. [132] to adopt the MIM strategy (Fig. 2.12). The input image is divided into multiple 16\*16 pixel image patches, and a part of the image patches is randomly masked. The method proposes a discrete variational autoencoder to

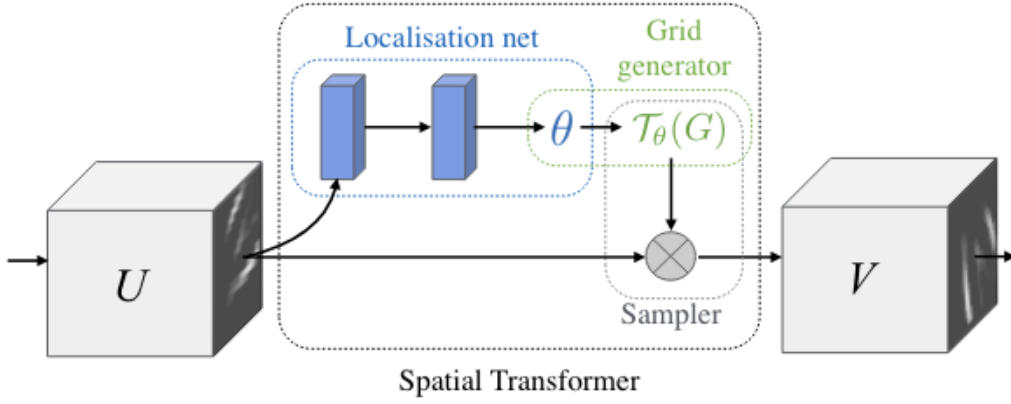


**Figure 2.13:** Architecture of Spatial Transformer Network [135]. The image intercepts from the original paper.

encode the masked image patches into a set of discrete IDs, and then predict the IDs of these masked image patches through ViT. Prior to this, they pre-trained a code book containing 3,000 categories and minimised the loss with the predictive visual tokens of the input image by ViT. Different from the complicated patch classification method in BEIT [132], SimMIM [133] uses a simpler framework, which is to directly regress and predict the original pixel RGB value. The mask patches are directly reconstructed by using a prediction head and increasing the image patches size to  $32 \times 32$ . SimMIM achieves 83.8% top-1 accuracy on ImageNet-1K [32]. Liu et al. [134] proposed a mixed MIM (MixMIM) strategy to create a mixed image by replacing image patches in two images, and reconstruct the original two images from the mixed image input. In order to avoid training errors caused by the significant appearance differences between the image patches of the two images, MixMIM uses a masked attention mechanism to prevent interaction between the two sets of image patches.

#### 2.2.4.2 Attention mechanism

Recent studies have shown that attention mechanisms have been widely used in CNN-based frameworks to reduce attention to irrelevant information and improve its ability to represent important regional features. Attention modules designed based on CNN usually have high independence and flexibility, which makes them easy to

apply to different CNN frameworks. Since hard attention selects a specific subset of the input image to focus on by sampling a discrete distribution, this sampling-based decision process is random and discontinuous. Therefore, hard attention is not differentiable and cannot be directly optimised by traditional gradient descent methods. It is usually trained by directly optimizing the target metric or reinforcement learning. In contrast, soft attention calculates continuous weights for different parts of the input image. The weight assignment process is continuous, and for any slight change in the weight, the output of the weighted sum will also change continuously accordingly. Due to the differentiability of soft attention, it can be easily integrated into deep learning models and trained by standard back-propagation algorithms. In response to the problem that CNN can only learn small-scale translation invariance, the Spatial Transformer Network proposed by Jaderber et al. [135] is used to explicitly learn various image transformations to make the model spatially invariant. As shown in the Fig. 2.13, Localisation net enlarges and centers the important regional features.  $\theta$  represents the spatial transformation parameters between the original image and the transformed image. In the Grid generator, the feature map is affine transformed, and bilinear interpolation is used to avoid mapping different positions before the transformation to the same coordinate when the affine transformation is rounded. Hu et al. [136] proposed Squeeze-and-Excitation (SE) Block to dynamically calibrate the channel response weights of the feature map by explicitly modeling the dependencies between image channels. In the Squeeze operation, each two-dimensional feature is compressed into a one-dimensional feature according to the spatial dimension, which represents the global distribution on the feature channel. Then the Excitation operation generates weights for each feature channel, and the weights are multiplied by the weights on the original features channel by channel to achieve reweighting of the original features. Woo et al. [137] combined the channel and spatial attention mechanisms and designed the Convolutional Block Attention Module (CBAM). CBAM includes the spatial attention module with global max pooling and global



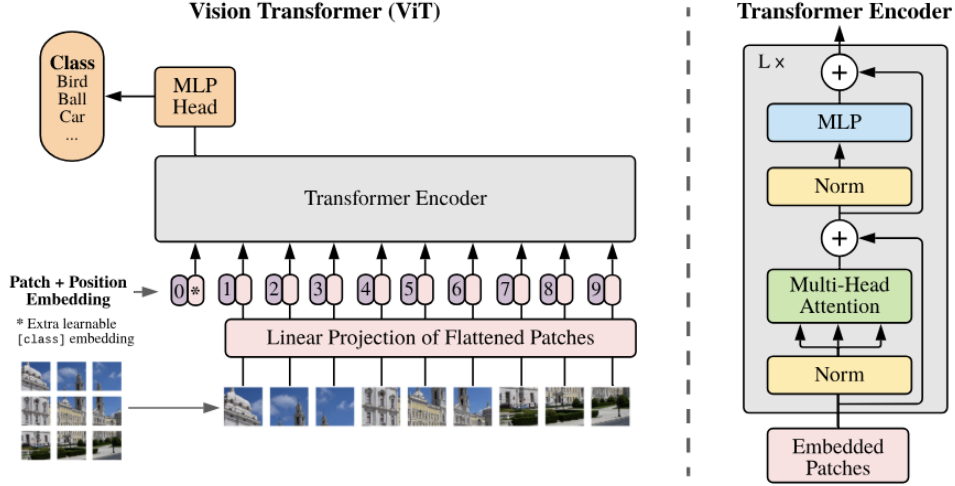


**Figure 2.14:** Architecture of Vision Transformer [131]. The image intercepts from the original paper.

average pooling applied to the width and height of the image, and the channel attention module based on SE Block.

As a special attention mechanism, the Self-Attention mechanism enables the network to have a global receptive field and capture the relationship between any part of the image. The most representative of these is the Transformer architecture, which has become the SOTA method in many machine translation tasks. Dosovitskiy et al. [131] proposed the Vision Transformer (ViT) to apply the Transformer architecture to computer vision tasks (as demonstrated in Fig. 2.15), and achieved the SOTA level after training on large-scale datasets such as ImageNet-21k [138]. ViT flattens the input image into fixed-size patches and each fixed-size patch is linearly embedded and added positional embedding. The resulting vector sequence is fed to the Transformer encoder. The Transformer encoder consists of alternating Multi-Head Self-Attention modules and Multi-Layer Perceptron modules connected by residual connection. Normalised vectors are linearly transformed into Query ( $Q$ ), Key ( $K$ ), and Value ( $V$ ) through different learnable weight matrices. Multi-Head Attention performs multiple linear transformations in parallel, each time using a different set of weight matrices  $W_i^Q$ ,  $W_i^K$  and  $W_i^V$  to generate multiple sets of Query, Key, and Value vectors. The purpose is to enable the model to

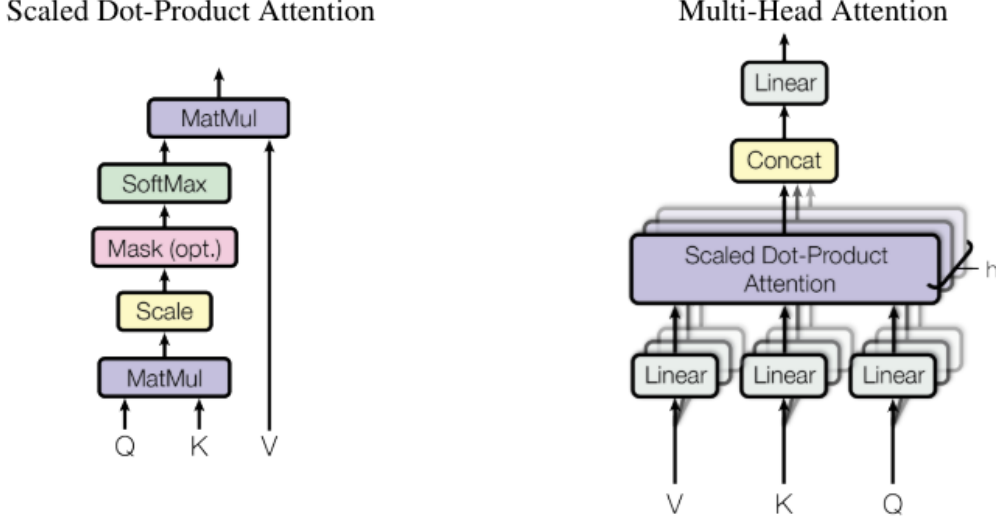


**Figure 2.15:** Architecture of Scaled Dot-Product Attention and Multi-Head Attention [139]. The image intercepts from the original paper.

capture information from different perspectives of the input data at the same time, increasing the model's representation capabilities.

The Scaled Dot-Product attention function for the same set of queries is as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (2.2)$$

$\frac{1}{\sqrt{d_k}}$  is the scaling factor. Finally, the outputs of all heads are concatenated and integrated through another linear layer. The formula can be expressed as:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (2.3)$$

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W^O \quad (2.4)$$

## 2.3 Summary

This chapter discusses the clinical and technical background of the thesis. The history of gastrointestinal endoscopy and the structure of the device is described in Section 2. This is followed by an introduction to Upper GI disease monitoring and Lower GI disease monitoring respectively, including a description of the disease as well as the

problems and challenges of the endoscopy process and subsequent surgical situations. This leads to the importance of using AI-based computer-aided examinations. Three classical computer vision tasks are reviewed in Section 2.2. Compared to traditional methods that rely heavily on manual initialisation and reliable prior knowledge, deep learning based methods are able to extract effective features more accurately and achieve superior results in different tasks. Literature on artefact detection is reviewed in Section 2.2.2.3. This thesis proposed an pruning-based mechanism to optimise artefact detection by comparing the profiles of each predicted instance with the pre-computed mean profile of that instance to achieve bounding box pruning. Recent advances in medical image classification especially disease in GI tract and computer vision are presented in Section 2.2.1.3 and Section 2.2.4 respectively. To address challenges such as scarcity of annotated data, the thesis proposed a self-supervised learning strategy using patch-level instance-group discrimination and penalisation of inter-class variation using additive angular margin. Based on the SOTA segmentation methods in GI endoscopy discussed in Section 2.2.3.3, the thesis focuses on video polyp segmentation that is more relevant to realistic endoscopy. This thesis presented a self-supervised pretext learning and temporal self-attention method with feature branching for real-time video polyp segmentation.

For this thesis, the findings in the literature review reveal critical gaps in current approaches, particularly around the limitations of manual initialization, the need for large annotated datasets, and the challenge of achieving real-time performance in video polyp segmentation. By identifying these challenges, this thesis proposes corresponding solutions. This review lays the foundation for the contribution of this thesis to address clinical and technical challenges in gastrointestinal endoscopy through innovative deep learning-based methods.

# 3

## Computer-aided image quality assessment in endoscopy

### Contents

---

|            |                                              |           |
|------------|----------------------------------------------|-----------|
| <b>3.1</b> | <b>Introduction</b>                          | <b>47</b> |
| <b>3.2</b> | <b>Methodology</b>                           | <b>48</b> |
| 3.2.1      | Artefact detection                           | 48        |
| 3.2.2      | Novel pruning method using instance profiles | 50        |
| 3.2.2.1    | Profile atlases                              | 51        |
| 3.2.2.2    | Detection box pruning                        | 52        |
| 3.2.3      | Quality score assessment                     | 52        |
| <b>3.3</b> | <b>Experiments</b>                           | <b>53</b> |
| 3.3.1      | Dataset                                      | 53        |
| 3.3.2      | Evaluation metrics                           | 54        |
| 3.3.3      | Experimental setup                           | 55        |
| <b>3.4</b> | <b>Results</b>                               | <b>56</b> |
| 3.4.1      | Quantitative results                         | 56        |
| 3.4.2      | Qualitative results                          | 56        |
| <b>3.5</b> | <b>Discussion &amp; Conclusion</b>           | <b>58</b> |

---

This chapter presents a deep learning-based image quality assessment system. It includes a novel bounding box pruning approach that can effectively improve artefact detection and provide high localisation scores of diverse artefact classes. Additionally, it incorporates an image quality scoring mechanism that relies on artefact detection.

**Originality:** The dataset used in this study comes from Endoscopy Artefact

Detection Challenge (EAD2020)[70, 140]. We train an EfficientDet architecture by minimising a focal loss, and compute the Bhattacharya distance between probability density of the pre-computed instance specific mean profiles of 7 artefact categories with that of the predicted bounding box profiles. Our results show that this novel approach is able to improve commonly used metrics such as mean average precision and intersection-over-union, by a large margin. This work was published in Medical Image Understanding and Analysis 2021.

- **Xu, Ziang**, Sharib Ali, Soumya Gupta, Numan Celik, and Jens Rittscher. "Improved artifact detection in endoscopy imaging through profile pruning." In Medical Image Understanding and Analysis: 25th Annual Conference, MIUA 2021, Oxford, United Kingdom, July 12–14, 2021, Proceedings 25, pp. 87-97. Springer International Publishing, 2021.

### 3.1 Introduction

Endoscopy is a widely used clinical procedure for assessing inflammation in hollow organs and performing subsequent treatments. Every endoscope typically consists of a light source and a camera connected to an external monitor to help clinicians navigate through the organ of interest in real-time. However, many factors such as lighting conditions, unstable motion, and organ structure result in presence of inevitable imaging artefacts that hinder the endoscopic surveillance and affects imaging data analysis. In clinical endoscopic surveillance of gastrointestinal organs, more than 60% of endoscopic video frames can be affected by artefacts [70]. A large number of artefacts can significantly affect the lesion detection algorithms, surface-based 3D reconstruction and other automated analysis methods [140, 141]. It is therefore extremely vital to find and locate these artefacts for a detailed endoscopic surveillance. While artefacts can be of various shapes, sizes and nature, accurate detection and precise localisation is difficult to achieve. We utilise deep learning-based object detectors to generate a list of potential artefacts, and then

exploit the nature of each artefactinstance to eliminate the false positives thereby enabling improved accuracy and localisation.

To improve on standard detection metrics, such as mean average precision (mAP) and intersection-over-union (IoU), while maintaining the test speed, we propose a pruning-based mechanism. We compare the intensity profiles of each predicted instance from the detector network with the pre-computed mean profile of that instance. To our knowledge this is the first work, that exploits the nature of each detected bounding box to prune detection by first computing probability density distribution of their center line grey-values and then comparing them with the computed mean profiles of 7 artefact instances. Here, we use EfficientDet [66], one of the state-of-the-art single-stage detector, with EfficientNet [37] backbone as our detector network. Our results demonstrate the effectiveness of our approach and shows improvement over both mAP and IoU metrics.

## 3.2 Methodology

We propose a comprehensive artefact detection method that comprises of: a) Bounding box prediction using EfficientNet with a bidirectional feature pyramid layer, b) NMS suppression to optimise the predictions, and c) proposed artefactprofile-based pruning of the artefactinstance boxes. Relevant details are presented in the following sections.

### 3.2.1 Artefact detection

We use EfficientNet [37] as the backbone for the EfficientDet [66] and stacked bi-directional feature pyramid network (BiFPN) layers to obtain high-level fusion features. The design of BiFPN layers comes from path aggregation network (PANet) [142]. It removes the node with only one input and connects the input and output nodes of the same level. The entire BiFPN layers are used as a base layer and are repeatedly connected three times as a feature network. It takes P3 to P7 features from the backbone network, and repeated two-way feature fusion. Both the bounding box and class prediction networks can share features and produce bounding box

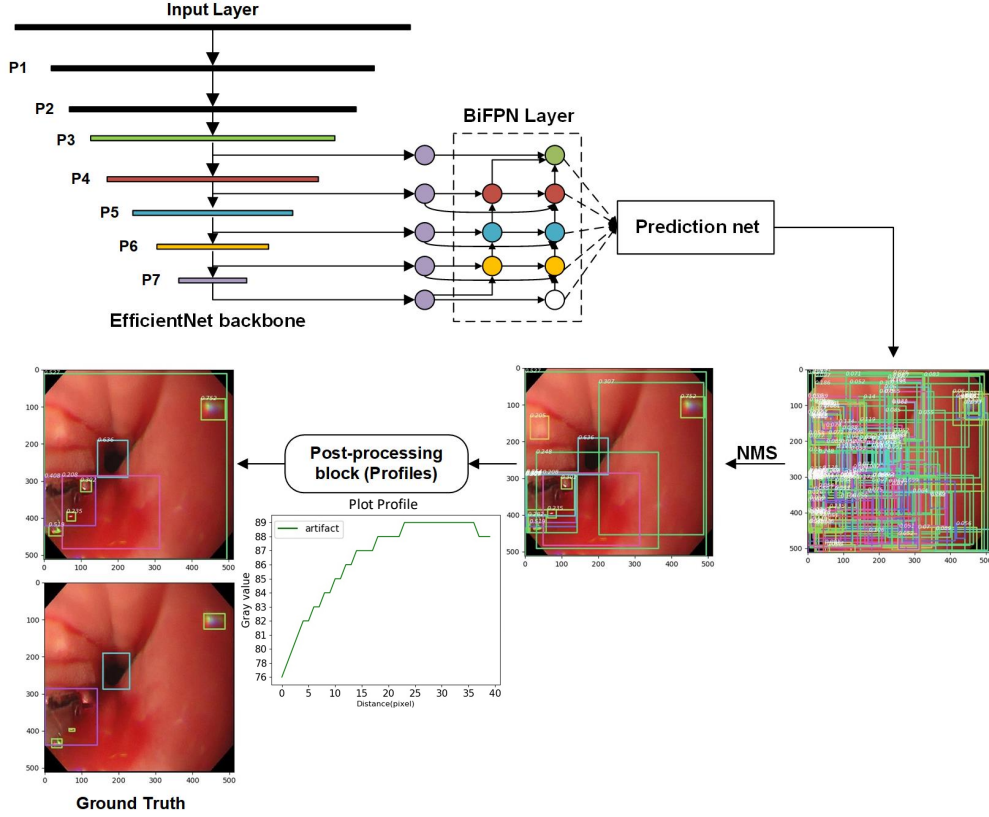


**Figure 3.1:** The architecture of proposed method. The overall process is divided into two parts, the artefact detection network and post-processing block based on artefactinstance profiles.

predictions and object classes, respectively. We modified the class prediction network for classifying 8 artefactclasses. Fig. 3.1 illustrates the architecture of EfficientDet.

Our goal is to minimise the total weighted loss  $\text{Loss}_{total}$  during training. Focal loss [65] with  $\alpha = 0.25$  and  $\gamma = 1.5$  and smooth L1 [69] loss  $L_{reg}$  are applied to the classification task and regression tasks, respectively. The regression loss is scaled by multiplication factor of 50 to balance the two loss functions,  $\lambda$  is set to 0.5. The resulting total loss  $L_T$  can be written as

$$L_T = (1 - \lambda)L_{cls} + 50 \times \lambda L_{reg} , \quad (3.1)$$

In addition, non-maximum suppression (NMS) [143] is used to optimise our results. As shown in Fig. 3.1, the direct prediction result of the model generates a large number of bounding boxes. The NMS selects the bounding box with the highest confidence in each category and calculates the IoU between the highest

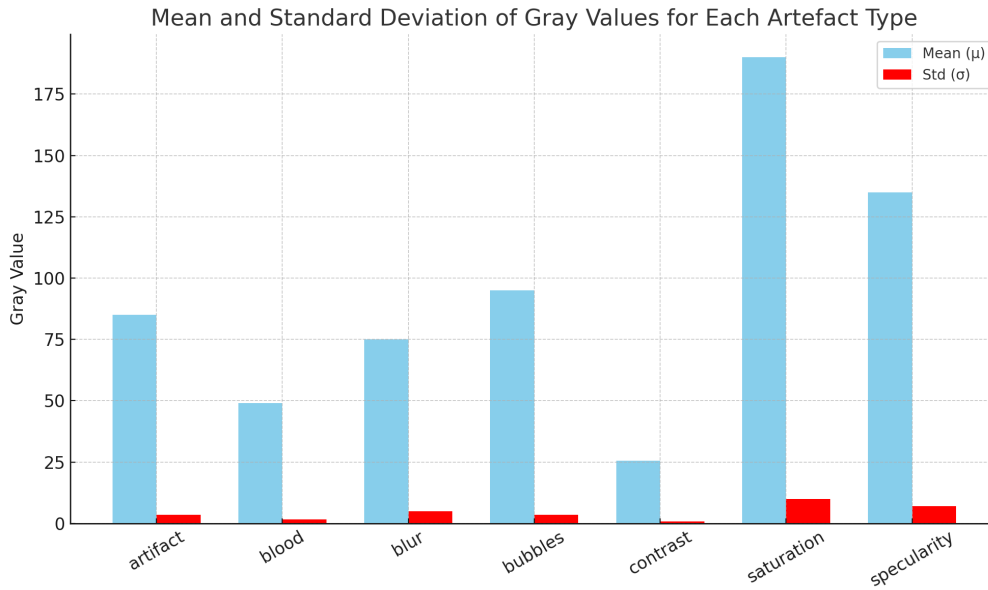


**Figure 3.2:** The mean ( $\mu$ ) and standard deviation ( $\sigma$ ) for center-line of 7 different artefact classes created from ground truth samples. The blue bars represent the average greyscale value for each artefact category. The red bars represent the corresponding standard deviation, indicating the spread of greyscale distribution within each category.

confidence bounding boxes and other bounding boxes one by one. If the score is greater than the set threshold, it is removed. It can be observed from the Fig. 3.1 that after NMS, the detection and localisation of artefacts has been greatly improved. However, compared with the ground truth, there are still erroneous predictions and multiple overlapping predictions.

### 3.2.2 Novel pruning method using instance profiles

Through detailed inspection of the dataset, we found that there are obvious differences in the profiles between different categories. It is specifically expressed as the probability density distribution (PDF) of grey values corresponding to the horizontal center-line pixels of the bounding boxes. For example, the characteristic of saturation profile is that the grey value quickly rises to the vertex (about 200) and then drops quickly. On the other hand, the line profile of contrast is extremely stable, and the grey value has been maintained at around 25. We also found that most of the artefacts are located in the center of the bounding boxes and the same artefacts have similar bounding boxes size (see Fig. 3.3). Additionally, since



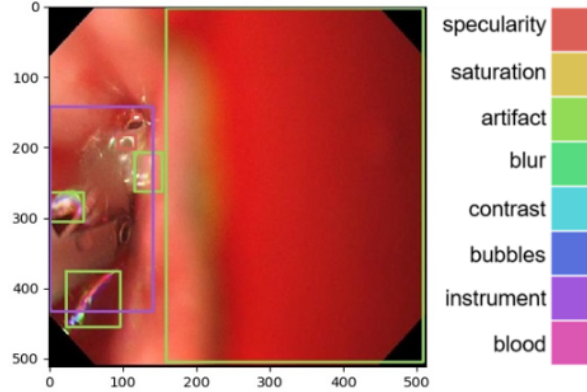


**Figure 3.3:** Examples of artefact locations in the ground truth image. The size of bounding boxes belonging to artefact class (green) are analogous and they are all on or close to the center line of bounding boxes.

instrument can be part of the imaging unlike other artefacts, so we have discarded instrument class in our work for profile pruning.

### 3.2.2.1 Profile atlases

We first computed the grey value distributions along the horizontal centre line of each bounding box in the ground truth annotations for all training samples. For each category, we first determined the representative centreline length by taking the median of the lengths across all its samples. To ensure uniformity across samples, we standardise the length of all profiles using the following strategy. If the centre line is longer than the representative length, we centre-crop it to the representative length. If the centre line is shorter than the representative length but longer than 20 pixels, we discard the profile, as interpolating these medium-length samples may distort the shape characteristics and introduce noise to the density distribution. If the centre line is shorter than 20 pixels, we pad both sides with zeros to reach the representative length. These very short profiles typically represent small artefacts where zero-padding has minimal impact and allows the retention of rare but informative samples. This strategy ensures all retained profiles within each category have the same length, balancing data quality and quantity. Once all profiles in a given category were standardised to the same length, we aggregated

all pixel values from all centreline profiles and calculated mean grey value for each artefact category, as shown in Fig. 3.2.

### 3.2.2.2 Detection box pruning

The prediction results obtained by the EfficientDet model include prediction categories, confidence scores and bounding boxes information. Profiles of the same style are extracted in each prediction result and are either shortened or elongated by adding zeros such that they could be compared with the profile atlases. Then the processed profiles and mean probability density distribution are compared to optimise the prediction results. We calculate the Bhattacharyya distance [144] between the two PDFs say  $p$  and  $q$  to reflect the difference between them. In statistics, the Bhattacharyya distance [145] is used to measure the similarity of two probability distributions and can be written as:

$$D_B(p, q) = \frac{1}{4} \ln \left( \frac{1}{4} \left( \frac{\sigma_p^2}{\sigma_q^2} + \frac{\sigma_q^2}{\sigma_p^2} + 2 \right) \right) + \frac{1}{4} \left( \frac{(\mu_p - \mu_q)^2}{\sigma_p^2 + \sigma_q^2} \right), \quad (3.2)$$

here  $\sigma_q^2$  and  $\mu_q$  represent the variance of the q-class distribution and the mean of the q-class distribution respectively.

In addition to the confidence score, we obtain another evaluation score, the Bhattacharyya distance score. Smaller Bhattacharyya distance means that the probability density distributions in two profiles are more similar. We keep the artefact bounding box predictions whose Bhattacharyya distance score is less than 4, and those greater than four are dropped out. The threshold of 4 was empirically determined based on validation set performance, where it was observed to provide a good trade-off between removing false positives (i.e., mismatched profile distributions) and retaining true artefact instances. This value effectively filters out unreliable predictions without significantly affecting detection sensitivity.

### 3.2.3 Quality score assessment

Based on the introduced artefact detection model and the subsequent instance profiles pruning method, we presented a frame quality scoring ( $QS$ ) assessment

which is proposed in [140]. The type of artefact, the percentage of the entire image area, and the location information of the artefact are considered separately. Different weights are assigned to each type according to the degree of visual interference. For example, compared with bubbles, artefacts have a greater visual impact on doctors. Therefore, the weight of artefacts is higher than that of bubbles. In addition, the area and location information of artefacts are also very important. artefacts located in the centre of the image will seriously block the doctor's view. Similarly, artefacts with a larger area will also reduce the image quality. The weighting scheme is as follows[140]:

- Class weight ( $W_{class}$ ): artefact(0.5), specularity (0.2), saturation (0.1), blur (0.05), contrast (0.05), bubbles (0.1).
- Location weight ( $W_{location}$ ): Center (0.5), left (0.25), right (0.25), top (0.25), bottom (0.25), top-left (0.125), top-right (0.125), bottom-left (0.125), bottom-right (0.125). The distribution of specific locations can be seen in the Fig. 3.4.
- Area weight ( $W_{area}$ ): The percentage of each artefact area and normal area in the whole image.

The proposed  $QS$  assessment relies on the accuracy of the pre-test task, instead it does not require additional data for training. The weights set are based on the recommendations of endoscopy experts. The final  $QS$  calculation is expressed as:

$$QS = 1 - \sum_B (\lambda_a W_{class} W_{area} + \lambda_l W_{class} W_{location}) \quad (3.3)$$

$B$  represents each bounding box where an artefact is detected.  $\lambda_a$  and  $\lambda_l$  weighs the contributions of area and location separately.

## 3.3 Experiments

### 3.3.1 Dataset

The dataset used is from the Endoscopy Artefact Detection Challenge (EAD2020) [70, 140]. There were eight different artefact categories, i.e. saturation, specularity,

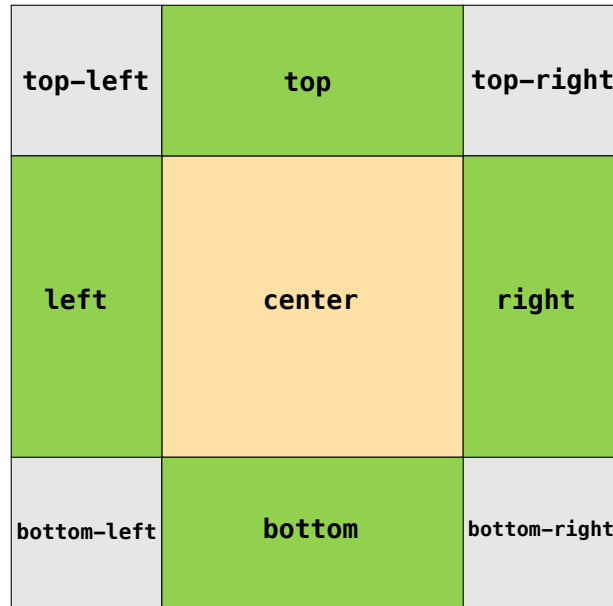


**Figure 3.4:** Location weight distribution in the quality scoring assessment.

bubbles, instrument, contrast, blood, blur and imaging artefacts. EAD2020 contains of 2200 single endoscopic frames and 232 sequential frames<sup>1</sup>. Each frame included an image frame and its corresponding bounding box annotations and category labels. We have used single endoscopic frames for training (80% of them are used for training set and the remaining 20% are used as validation set), and sequential frames are used as the out-of-sample test set. Table 4.1 provides the total number of artefacts present in the training data which split by classes.

### 3.3.2 Evaluation metrics

In this work, we have computed widely used detection metrics described below:

- Intersection-over-union (IoU): This metric quantifies the area overlap between two spatial regions using the intersection-over-union between reference or ground-truth (denoted R) and predicted bounding boxes and segmentations

<sup>1</sup><http://dx.doi.org/10.17632/c7fjbxcgj9.3>

**Table 3.1:** Statistics of the training data.

| Class       | Number of instances |
|-------------|---------------------|
| Specularity | 8596                |
| artefacts   | 7688                |
| Bubbles     | 4627                |
| Contrast    | 1590                |
| Saturation  | 1184                |
| Blur        | 671                 |
| Instrument  | 446                 |
| Blood       | 446                 |
| Total       | 25248               |

(denoted  $S$ ).

$$IoU(R, S) = \frac{R \cap S}{R \cup S}. \quad (3.4)$$

- Mean average precision (mAP): This metric measures the ability to accurately retrieve all ground truth bounding boxes. mAP calculates the average precision (AP) of each category based on different IoU thresholds, and then the average of all categories of AP are estimated. AP refers to the calculation of the area under curve (AUC) of the precision-recall curve after obtaining the maximum Precision value.

### 3.3.3 Experimental setup

We implemented EfficientDet using PyTorch <sup>2</sup>. In the training process, we first resized the training data to 512×512. We used Adam optimiser with a weight decay of 1e-4 and learning rate of 1e-3 for training the model. In the loss function, we used  $Loss_{total}$  to include Focal loss with  $\alpha = 0.25$  and  $\gamma = 1.5$  and a smooth  $L1$ -loss regularisation. The model was trained for 1500 epochs on NVIDIA Quadro RTX 6000. The results of network prediction were optimised by NMS with set threshold of 0.5. We additionally set a confidence threshold of 0.2 to further eliminate false positive predictions. The results were then fed to the post-processing block for further pruning and optimization. As described in the method section, we got our final result by calculating the distance and discarding all scores greater than

---

<sup>2</sup><https://pytorch.org>

**Table 3.2:** Experimental results on testing set. Each model has 16 million parameters.

| Test Model              | mAP <sub>25</sub> | mAP <sub>50</sub> | mAP <sub>overall</sub> | IoU           |
|-------------------------|-------------------|-------------------|------------------------|---------------|
| EfficientDet-d0         | 30.67%            | 20.93%            | 20.29%                 | 5.17%         |
| EfficientDet-d0-NMS     | 29.56%            | 19.77%            | 19.93%                 | 11.60%        |
| YOLOv3 [146]            | <b>32.19%</b>     | 18.47%            | 17.17%                 | 16.35%        |
| RetinaNet [65]          | 17.64%            | 6.45%             | 8.08%                  | 10.00%        |
| DuyHUYNH [71]           | 28.63%            | 15.52%            | 15.46%                 | 16.96%        |
| EfficientDet-d0-NMS-PDF | 31.69%            | <b>19.98%</b>     | <b>21.33%</b>          | <b>19.01%</b> |

4, which is empirically set based on our experimental validation of improvement. For the  $QS$  assessment,  $\lambda_a = \lambda_l = 0.5$  is set.

## 3.4 Results

### 3.4.1 Quantitative results

Table 4.2 summarises the result of artefacts detection on test set. It can be seen that EfficientDet-d0 model got overall mAP of 20.29% and an IoU of 5.17%. This means the model over predicted the class instances with poor localisation. However, after the NMS, overall mAP dropped to 19.93% with increased IoU by 6.43% (i.e., IoU of 11.60%). However, by using our proposed elimination technique, (EfficientDet-d0-NMS-PDF), where we compare distance between the predicted PDF and the mean PDF, the results are improved. It can be observed that with technique we achieve an overall mAP of 21.33% and an IoU of 19.01% (an increase by nearly 15%). Similarly, we observed improvements in mAPs at other IoU thresholds (e.g, mAP<sub>25</sub> of 31.69%, mAP<sub>50</sub> of 19.98%) with EfficientDet-d0-NMS. Compared to YOLOv3 [146] and DuyHUYNH [71], our approach demonstrates substantial improvements in both overall mAP and IoU metrics.

### 3.4.2 Qualitative results

As shown in Fig. 3.5, the results after NMS processing have been significantly improved, but there is large number of wrong bounding box predictions. Our PDF-based post-processing method further optimises the results and reduces the incorrect predicted bounding boxes. In the original prediction of the model, there is a large

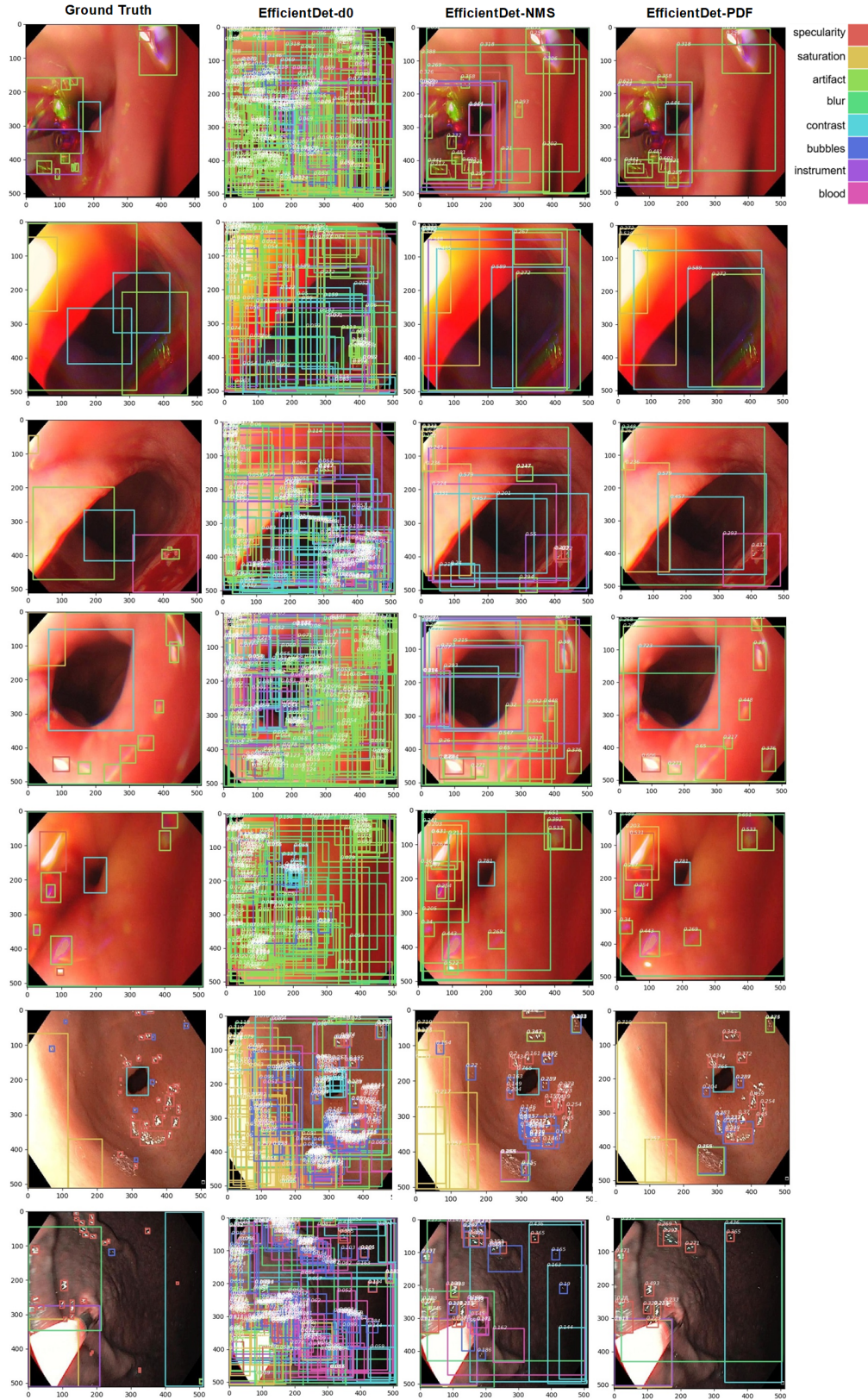


**Figure 3.5:** Qualitative Results of EfficientDet-d0, EfficientDet-NMS and EfficientDet-NMS-PDF. Different colored bounding boxes represent different classes.

number of erroneous results between different artefacts, especially in specularity and artefact, as well as prediction bounding boxes of different sizes for the same artefact area, such as contrast. Our proposed method prunes these predictions by calculating the Bhattacharyya distance-based score. From the Qualitative Results point of view, our results have fewer and more accurate annotation bounding boxes which can be observed to be closer to the ground truth annotations.

Fig. 3.6 shows results of the *QS* assessment. In the first row, because bubbles have a low weight and occupy small areas, the image quality score of the first row is high, reaching 0.953. The second row is mainly composed of artefacts, which are large in area and account for 49.2% of the entire image. artefacts have a greater weight because they have a more serious impact on line of sight occlusion. In particular, the location of the artefacts in the image is concentrated in the center of the image, so the quality score is only 0.252. The third row contains a variety of artefacts, including specularity, saturation and contrast. Although the weight of specularity is high and the location of specularity is concentrated in the center of the image, its area is very small (2.24% of the total image area), so it has little impact on the quality score. The area of the saturation in the figure is about 23%, even if it is on the left side of the image, it still has a certain impact. The final quality score is 0.529.

### 3.5 Discussion & Conclusion

In this chapter, we proposed a novel method of eliminating wrong predictions based on distances between the horizontal line profile of the predicted bounding boxes with that of the pre-computed atlas line profiles. The proposed method achieves the purpose of bounding box pruning by calculating the Bhattacharyya distance between instances. State-of-the-art EfficientDet is trained for artefact detection tasks. The experimental results show that our method effectively improves the mean average precision and intersection-over-union. Compared with the widely used non-maximum suppression technique, our method has a higher metric scores. It can be also seen from our qualitative experiment that the results of the proposed PDF-based



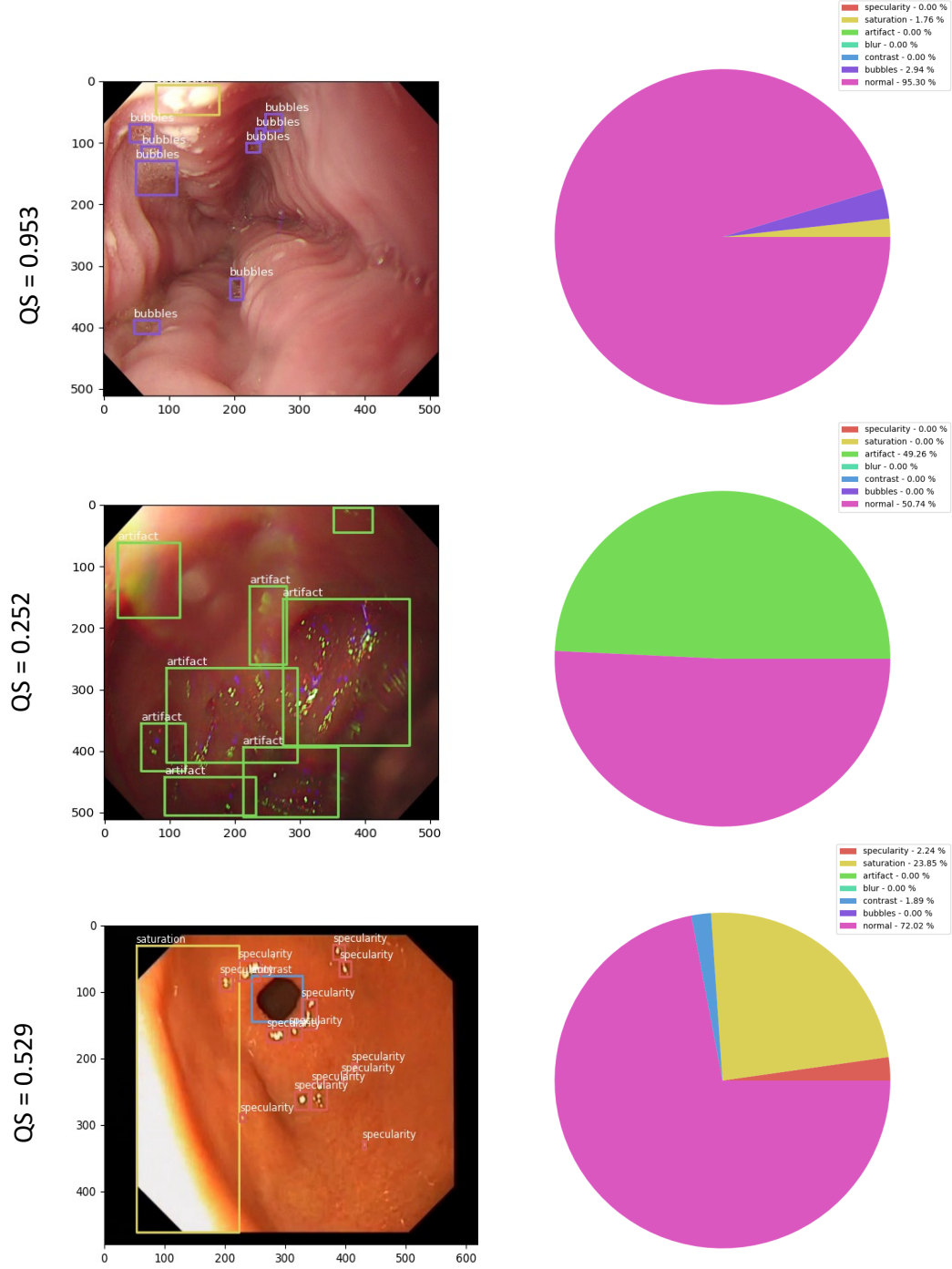


**Figure 3.6:** *QS* assessment focus on artefact detection, location and area [140]. The left image shows the image with detection boxes in the artefact detection task. The right image shows the corresponding area percentage. On the far right are their quality scores.

method are closer to the ground truth bounding box annotations. Additionally, the flexibility of the proposed pruning method, which is model-agnostic, allows it to be applied to any deep learning-based object detection framework.

The quality score assessment mechanism introduced in this chapter can serve as an essential preprocessing step for other algorithms used in endoscopy image analysis. In clinical settings, low-quality images (e.g., images with artefacts or poor illumination) can compromise the accuracy of downstream diagnostic tasks, such as lesion detection or segmentation. By incorporating the proposed quality evaluation technique, frames or images with suboptimal quality can either be flagged for review or excluded from further processing. In future developments, the quality score could become a filtering stage that precedes more advanced diagnostic models. For instance, a segmentation model could first rely on this quality metric to ensure that it is processing only high-quality, artefact-free images, which could reduce errors and increase the overall performance of the system. Moreover, by automatically assigning quality scores, the system could prioritise high-quality images for immediate processing, while lower-quality ones could undergo enhancement techniques before feeding into more advanced algorithms.

The proposed quality score assessment can evaluate the score of each frame and provide image quality feedback to clinicians. Real-time feedback during endoscopic procedures, based on the algorithm’s analysis of the image quality and detection accuracy, could be invaluable. For instance, when the system identifies low-quality frames (e.g., frames with artefacts or blurred regions), it could provide real-time prompts to the clinician to adjust the scope’s position, improve lighting, or take corrective actions. This feedback loop could ensure that high-quality data is captured throughout the procedure, leading to better outcomes.

In the future, as the system continues to improve and refine its pruning techniques, it could also offer more detailed suggestions, such as recommending re-examination of specific areas that had suboptimal image quality or ambiguous detections. Over time, this could reduce the rate of missed diagnoses and improve the thoroughness of endoscopic evaluations. Additionally, by providing detailed feedback and quality

assessments after a procedure, clinicians could use the information for training purposes, helping to refine their techniques and improve their own diagnostic accuracy over time.

Endoscopic procedures often involve video sequences rather than single frames, so aggregating information from multiple frames is crucial for developing a comprehensive understanding of the anatomical structures and potential artefacts. One method to achieve this is through temporal aggregation, where the predictions from multiple consecutive frames are analysed together. For example, if a certain frame's detection is uncertain (i.e., the bounding box is pruned due to a high Bhattacharya distance), but surrounding frames show a strong and consistent prediction, the system could aggregate this information to retain the bounding box in the uncertain frame, thereby improving the overall detection consistency.

Another approach could involve using temporal consistency checks, where the bounding box predictions are smoothed across adjacent frames to reduce jittering or sudden changes in detection results. This method would take into account the spatial and temporal correlations between frames, ensuring that artefacts or regions of interest are consistently tracked throughout the video. A post-processing step could involve summarizing the quality scores across the entire video sequence, identifying regions with lower confidence or poor quality that might warrant re-examination or further attention.

In summary, this chapter presents a novel approach to improving artefact detection and bounding box pruning. Approaches to image quality assessment based on artefact detection hold great potential for research and application in the clinical setting by providing feedback to clinicians and acting as a precursor algorithm to other disease detection algorithms.

# 4

## Computer-aided risk assessment in ulcerative colitis

### Contents

---

|            |                                                                                                                    |           |
|------------|--------------------------------------------------------------------------------------------------------------------|-----------|
| <b>4.1</b> | <b>Supervised additive angular margin loss and model scaling network for optimised colitis scoring . . . . .</b>   | <b>64</b> |
| 4.1.1      | Introduction . . . . .                                                                                             | 64        |
| 4.1.2      | Methodology . . . . .                                                                                              | 66        |
| 4.1.2.1    | Model scaling network . . . . .                                                                                    | 67        |
| 4.1.2.2    | ArcFace loss . . . . .                                                                                             | 67        |
| 4.1.3      | Experiments . . . . .                                                                                              | 68        |
| 4.1.3.1    | Dataset . . . . .                                                                                                  | 68        |
| 4.1.3.2    | Evaluation metrics and experimental setup . .                                                                      | 69        |
| 4.1.4      | Results . . . . .                                                                                                  | 69        |
| 4.1.5      | Contribution and conclusion . . . . .                                                                              | 70        |
| <b>4.2</b> | <b>Self-supervised learning with composite pretext-class discrimination for improved colitis scoring . . . . .</b> | <b>71</b> |
| 4.2.1      | Introduction . . . . .                                                                                             | 71        |
| 4.2.2      | Methodology . . . . .                                                                                              | 74        |
| 4.2.2.1    | Feature extraction (FE) block . . . . .                                                                            | 76        |
| 4.2.2.2    | Contrastive loss estimation with margin (CEM) block . . . . .                                                      | 77        |
| 4.2.2.3    | $k$ -means feature grouping . . . . .                                                                              | 79        |
| 4.2.2.4    | Cross-level discrimination at image and patch-levels . . . . .                                                     | 79        |
| 4.2.2.5    | Proposed loss . . . . .                                                                                            | 80        |
| 4.2.3      | Experiments . . . . .                                                                                              | 81        |
| 4.2.3.1    | Dataset . . . . .                                                                                                  | 81        |
| 4.2.3.2    | Evaluation metrics . . . . .                                                                                       | 82        |
| 4.2.3.3    | Implementation details . . . . .                                                                                   | 82        |

|         |                                                 |    |
|---------|-------------------------------------------------|----|
| 4.2.4   | Results . . . . .                               | 83 |
| 4.2.4.1 | Comparison for UC classification task . . . . . | 84 |
| 4.2.4.2 | Computation time . . . . .                      | 84 |
| 4.2.4.3 | Generalisation . . . . .                        | 84 |
| 4.2.4.4 | Ablation studies . . . . .                      | 85 |
| 4.2.4.5 | Qualitative Analysis . . . . .                  | 89 |
| 4.2.5   | Contribution and conclusion . . . . .           | 91 |

---

This chapter presents methods for ulcerative colitis classification, which are divided into supervised learning and self-supervised learning. The supervised learning method improves the classification effect by improving the additive angular margin loss function. The self-supervised learning method achieves state-of-the-art accuracy by pre-training the model on a uniquely designed pretext task and fine-tuning it in the downstream task.

**Originality:** The datasets used are from the public datasets HyperKvasir [3], LIMUC [147] and an in-house dataset (collected from John Radcliffe Hospital, Oxford and labeled by Prof. James East). In Section 4.1, we first propose a 3-way Mayo endoscopic subscore (MES) classification by using an additive angular margin loss function in addition to cross-entropy loss and model scaling network. In Section 4.2, we propose to explore patch-level instance-group discrimination and penalisation of inter-class variation using additive angular margin within the cosine similarity metrics. Our novel approach enables models to learn to cluster similar representations, thereby improving their ability to provide better separation between different classes. The supervised learning method was published in IEEE 19th International Symposium on Biomedical Imaging (ISBI). The self-supervised learning method was first published in International Workshop on Machine Learning in Medical Imaging held in conjunction with Medical Image Computing and Computer Assisted Intervention 2022 (MICCAI), and then we extended our work and published it in IEEE Transactions on Medical Imaging.

- **Xu, Ziang**, Sharib Ali, James East, and Jens Rittscher. "Additive angular margin loss and model scaling network for optimised colitis scoring." In 2022

IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1-5. IEEE, 2022.

- **Xu, Ziang**, Sharib Ali, Soumya Gupta, Simon Leedham, James E. East, and Jens Rittscher. "Patch-level instance-group discrimination with pretext-invariant learning for colitis scoring." In International Workshop on Machine Learning in Medical Imaging, pp. 101-110. Cham: Springer Nature Switzerland, 2022.
- **Xu, Ziang**, Jens Rittscher, and Sharib Ali. "SSL-CPCD: Self-supervised learning with composite pretext-class discrimination for improved generalisability in endoscopic image analysis." IEEE Transactions on Medical Imaging (2024).

## 4.1 Supervised additive angular margin loss and model scaling network for optimised colitis scoring

### 4.1.1 Introduction

The human gastrointestinal system is very fragile and vulnerable to viruses. Gastric cancer and colorectal cancer cause about 1.8 million deaths each year accounting for more than 12% of the global mortality rate [1]. Colonoscopy is a gold standard clinical procedure for early screening of cancer. Studies have shown that patients with inflammatory bowel disease (IBD) have an increased lifetime risk of developing colorectal cancer (CRC) [12]. Monitoring and assessing disease progression and response to drug therapy over time is thus critical. Compared with Crohn's disease (CD), ulcerative colitis (UC) mostly occurs in colon and rectum and has more focused features. Grading of UC can extend over several different scoring protocols, e.g. Mayo Endoscopic Score (MES) [14] that is characterised and graded based on rectal bleeding, vascular pattern, mucosal topology etc. UC can cause significant damage to the rectal mucosa that can lead to rectal dysplastic diseases and further to cancer. Early detection of UC is important as UC can be treated at an early

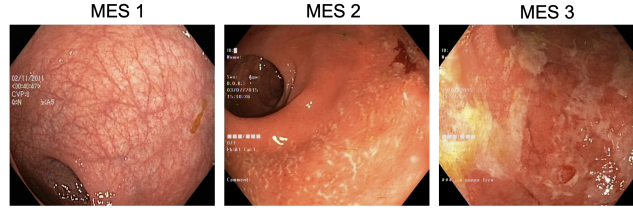


**Figure 4.1:** Mayo endoscopic scoring representative images

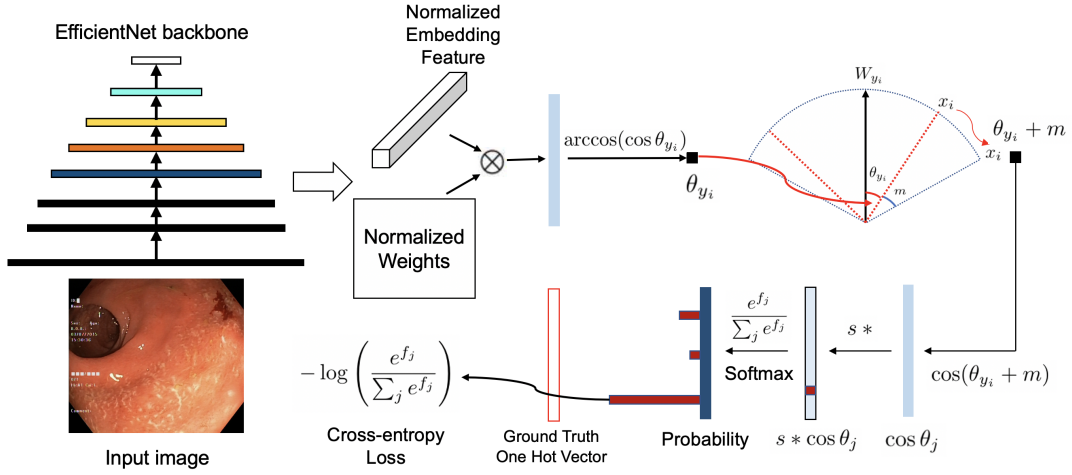


**Figure 4.2:** Training of a model scaling network for improved colitis classification with ArcFace loss [149].

stage, however, the interpretation and scoring of UC is highly experience-based. Although IBD specialists can perform follow-up scoring, it has been reported that there is a significant discrepancy between the colonoscopy and the follow-up scores for UC [148]. Thus, deep learning-based methods can help to minimise operator dependence observed in complex IBD scoring.

In recent years, deep learning has been widely used in various medical image analysis tasks. In the UC severity classification task, most groups have used deep learning-based convolutional neural networks (CNNs) [38, 57, 150] to build the UC scoring system. We propose a novel classification technique for UC grading based on MES scoring protocol. We use deep learning-based model scaling network combined with additive angular margin loss function to achieve three classes (MES 1, MES 2 and MES 3) classification task. In contrast, most of the previous works are based on binary UC scoring. Compared with binary UC scoring, multi-

class UC scoring will enable a more continuous scoring strategy for improved risk assessment. The phenotypic changes that are costumed in MES scoring for each class is defined below [151]:

- i. MES 1 (mild UC) is characterised by erythema, decreased vascular pattern, mild friability (see Fig. 4.1, left)
- ii. MES 2 (moderate UC) is characterised by marked erythema, absence of vascular pattern, friability, erosions (see Fig. 4.1, middle)
- iii. MES 3 (severe UC) is characterised by spontaneous bleeding and ulceration (see Fig. 4.1, right)

Most deep learning-based approaches for UC classification adopt the standard cross-entropy loss [38, 57]. However, the UC grading task is characterised by subtle visual differences between categories, which makes it difficult for cross-entropy to effectively separate class boundaries. To address this, we adopt the ArcFace loss instead. ArcFace introduces an angular margin between classes, thereby enhancing inter-class separability and intra-class compactness. This leads to a more discriminative feature embedding space, improving classification robustness and precision in challenging endoscopic conditions.

Our contribution includes - 1) identifying a model scaling approach that improves the state-of-the-art classification networks such as ResNet and DenseNet, 2) investigation on integration of additive angular margin loss function that can improve classification accuracy largely when optimal parameters for margin and scale are setup, and 3) three class MES scoring technique unlike previous binary approaches.

### 4.1.2 Methodology

The overall approach for our model scaling with additive angular margin loss for colitis scoring is shown in Fig. 4.2. Below we describe different components of our approach.



#### 4.1.2.1 Model scaling network

The model scaling strategy applied in EfficientNet [37] can effectively balance the network depth ( $d$ ), width ( $w$ ) and resolution ( $r$ ) to achieve better performance. The composite coefficient ( $\phi$ ) is defined as the strategy to complete the unified scaling of all dimensions of depth ( $d = \alpha^\phi$ ), width ( $w = \beta^\phi$ ) and resolution ( $r = \gamma^\phi$ ). By fixing the basic network layer ( $\alpha = \beta = \gamma = \text{constant}$ ), model scaling simplifies resource constraints, but it still has a relatively large search space. In order to further reduce the design space, the method restricts all layers to be uniformly scaled with a constant scale to achieve the goal of maximizing model accuracy. Specifically,  $d = \alpha^\phi$  controls how many times each mobile inverted bottleneck convolution (MBConv) [152] block is repeated;  $w = \beta^\phi$  determines the number of channels in each block; and  $r = \gamma^\phi$  defines the input image resolution. By incrementally increasing the value of  $\phi$ , EfficientNet expands the model uniformly across all three dimensions to achieve better performance while maintaining computational efficiency. The baseline network in EfficientNet-b0 ( $\phi = 1$ ) is obtained from a multi-objective neural network structure search. EfficientNet-b0 consists of 16 MBConv modules with varying kernel sizes (e.g.,  $3 \times 3$  and  $5 \times 5$ ), strides, and expansion ratios. Each block contains 32 channels, and the input image resolution is fixed at  $224 \times 224$ .

#### 4.1.2.2 ArcFace loss

ArcFace [149] introduces an additive angular margin penalty based on softmax, which is a geometrically interpretable loss function. We modified the layer of our classification network after the last convolutional layer, i.e. fully connected feature layer, to obtain the feature embedding which is then fed to the ArcFace loss function. The weights and features are L2 normalised such that the learned embeddings are distributed on a unit hyper-sphere. An angle  $\theta_{y_i}$  is computed as the product between normalised feature embedding and weights, where  $y_i$  represent logits  $i \in [0, 1, 2, \dots, N]$ . An additive angular margin penalty ‘ $m$ ’ is added to the angular projection of feature embedding ( $\theta_j = \theta_{y_i} + m$ ) such that the decision boundary between classes in the angular space is maximised pushing inter-class

samples and pulling intra-class samples. Finally, all logits are scaled by a factor ‘s’ and softmax is used to obtain the final class probabilities defined as:

$$L_{ArcFace} = -\log \frac{e^{s \cos(\theta_{y_i} + m)}}{e^{s \cos(\theta_{y_i} + m)} + \sum_{j=1, j \neq y_i}^N e^{s \cos(\theta_j)}}, \quad (4.1)$$

where  $\theta_j$  is the angle between the embedding feature  $x_i$  of the  $i$ -th sample and the  $j$ -th ground truth (target) weight  $W_j$  which behaves as a centre for each class. Here,  $y_i$  represents the corresponding class label of  $x_i$ ,  $N$  is the total class number, and  $s$  and  $m$  are the parameter of feature re-scale and angular margin, respectively. Cross entropy loss is then minimised between ground truth and label probabilities (see Fig. 4.2). It has been empirically shown [149] that setting  $m = 0.5$  and  $s = 32$  provides optimal results. However, we have searched these values for MES-scoring.

### 4.1.3 Experiments

#### 4.1.3.1 Dataset

We have used HyperKvasir [3] dataset that provides six different UC categories, including Mayo scores (1,2 and 3) and three additional scoring levels categorising into scores 0-1, 1-2 and 2-3, i.e. ambiguous frames. In total there are 851 images for the UC dataset. We have revised this dataset with an input from an expert colonoscopist (> 20 years of experience) to reduce the data ambiguity. The final data is categorised into three distinct grades (i.e., 1 = mild, 2 = moderate and 3 = severe). For our experiments, we randomly selected 80% of data for training (681 images), 10% data for validation (85 images) and 10% for test (85 images). The complete description of the revised dataset and codes will be made publicly available.

**Table 4.1:** Experimental results on test set for classification of MES scores (1, 2 and 3). Two best results are presented in bold.

| Model                           | Top 1 accuracy | Top 2 accuracy | F1 score      | Computational complexity | Time (ms) |
|---------------------------------|----------------|----------------|---------------|--------------------------|-----------|
| DenseNet-161 (26.7M)            | 60.33%         | 91.79%         | 62.36%        | 30.1M                    | 9         |
| ResNet-152 (60.4M)              | 63.58%         | 89.79%         | 64.91%        | 60.3M                    | 11        |
| EfficientNet-b0 (5.3M)          | 70.83%         | 91.75%         | 71.59%        | 6M                       | 7.9       |
| EfficientNet-b1 (7.8M)          | 71.25%         | 91.71%         | 71.86%        | 8M                       | 8.3       |
| EfficientNet-b3 (12M)           | 71.87%         | 90.58%         | 73.46%        | 12M                      | 9.1       |
| EfficientNet-b4 (19M)           | 71.83%         | 90.62%         | 71.45%        | 19M                      | 9.9       |
| EfficientNet-b7 (66M)           | <b>72.95%</b>  | 90.58 %        | <b>75.00%</b> | 66M                      | 12        |
| ResNet-152 + $L_{ArcFace}$      | 67.89%         | <b>92.27%</b>  | 68.56%        | 60.3M                    | 11        |
| EfficientNet-b7 + $L_{ArcFace}$ | <b>75.06%</b>  | <b>93.68%</b>  | <b>76.15%</b> | 66M                      | 12        |

#### 4.1.3.2 Evaluation metrics and experimental setup

We implemented our proposed method using PyTorch<sup>1</sup>. Training and validation data were resized to  $224 \times 224$ . SGD optimiser with a weight decay of  $4e^{-4}$ , momentum of 0.9 and learning rate of  $1e^{-3}$  was used for training the model. We also set the learning rate decay strategy with a learning rate decay of 0.8 times per 1000 epochs. Training was conducted for 10,000 iteration on NVIDIA Quadro RTX 6000 with a batch size of 64. We have used standard F1 score ( $= \frac{tp}{tp+fp}$ , tp: true positive, fp: false positive) and top-k accuracy (percentage of samples predicted correctly) for our 3-way classification task of MES-scoring for UC. All models were trained on the same training dataset using the same training strategy.

#### 4.1.4 Results

From Table 4.1, baseline networks ResNet-152 and DenseNet-161 obtained 60.33% and 63.58%, respectively, on top-1 accuracy. Compared to the baseline network, EfficientNet showed improved accuracies with EfficientNet-b0, which is the least computational complexity, achieving 70.83% of top-1 accuracy and 71.59% of F1 score (an increase by 7.25% and 6.68% compared to ResNet-152). As the EfficientNet model continues to be enlarged (i.e.  $\phi > 1$ ), the accuracy of the model continues to improve. With the EfficientNet-b7 we obtained the best result of 72.95% for top-1 accuracy and 75.00% for F1 score. Our final modification with ArcFace loss resulted in further improvements. This increase in classification metrics was also observed for the ResNet-152 showing an increment of 4.31%, 2.48% and 3.65% on top-1 accuracy, top-2 accuracy and F1 score, respectively. However, our choice of network, i.e. EfficientNet-b7 together with ArcFace, outperformed all the other classification models both with and without ArcFace resulting in an accuracy of 75.06% for top-1 accuracy, 93.68% for top-2 accuracy and 76.15% for F1 score.

---

<sup>1</sup><https://pytorch.org>

**Table 4.2:** Angular and scale search for our ArcFace loss  $L_{ArcFace}$  on test set

| Model                                | Accuracy      |               | F1-score      |
|--------------------------------------|---------------|---------------|---------------|
|                                      | Top1          | Top2          |               |
| EfficientNet-b7-Arc<br>(m=0.3, s=2)  | 73.62%        | 91.97%        | 74.02%        |
| EfficientNet-b7-Arc<br>(m=0.3, s=6)  | 73.98%        | 92.76%        | 74.87%        |
| EfficientNet-b7-Arc<br>(m=0.3, s=10) | 72.90%        | 91.37%        | 73.10%        |
| EfficientNet-b7-Arc<br>(m=0.5, s=2)  | 74.77%        | 92.10%        | 75.03%        |
| EfficientNet-b7-Arc<br>(m=0.5, s=6)  | <b>75.06%</b> | <b>93.68%</b> | <b>76.15%</b> |
| EfficientNet-b7-Arc<br>(m=0.5, s=10) | 73.48%        | 91.79%        | 74.33%        |
| EfficientNet-b7-Arc<br>(m=0.7, s=2)  | 74.56%        | 93.43%        | 75.16%        |
| EfficientNet-b7-Arc<br>(m=0.7, s=6)  | 74.59%        | 92.39%        | 75.09%        |
| EfficientNet-b7-Arc<br>(m=0.7, s=10) | 73.67%        | 91.88%        | 74.79%        |

**Ablation study** The setting of feature re-scaling ( $s$ ) and angular margin ( $m$ ) parameters in ArcFace loss affects the model performance. Therefore, we conducted an extensive experiment to investigate the performance of the ArcFace loss under different parameter settings. We set  $s = 2, 6, 10$ ,  $m = 0.3, 0.5, 0.7$  as our experimental parameters for scale and angular margin. From Table 4.2, we can see that ArcFace loss has a significant performance drop (top-1 accuracy from 75.06% to 73.48%) when  $s$  increases from 6 to 10. Similarly, if the value of  $s$  is reduced to 2, the performance of the model still decreases slightly. When  $m = 0.5$ , the performance of the model is slightly higher (about 1%) than other values. Finally, for  $m = 0.5$  and  $s = 6$  the chosen EfficientNet-b7 yields the best classification accuracies.

#### 4.1.5 Contribution and conclusion

In this chapter, we proposed to use a model scaling network to deal with classification task of ulcerative colitis which is a known complex mucosal classification task. To improve the classification performance of the model we integrated ArcFace loss

that accounted for highly similar features in colonic images. Our experiments showed improvements over various SOTA classification approaches which evidently concludes that our approach is able to generate more distinct feature embedding and the incorporation of ArcFace allowed to increase the inter-class distance between different categories, thereby improving the classification accuracies.

## 4.2 Self-supervised learning with composite pretext-class discrimination for improved colitis scoring

### 4.2.1 Introduction

Recent advances in data-driven approaches are mostly based on convolutional neural networks (CNNs) and have gained interest due to their ability to surpass traditional machine-learning approaches. CNNs have been widely used for multiple tasks and different imaging modalities, including computed tomography (CT) [153], X-ray [154], magnetic resonance imaging (MRI) [155] and endoscopy [156].

Supervised learning-based approaches in machine learning (ML) are data-voracious. Performance of fully supervised methods usually suffers on smaller datasets and out-of-distribution datasets, which can be because supervised learning only incentivises learning those features that are relevant to predicting frequent classes of known samples. [157]. Obtaining labelled data is a significant hurdle for medical image analysis as it requires clinical expertise. Additionally, it accounts for the risk of human bias proportional to the sample size [158]. Data curation challenges are thus harder to tackle, leading only to sub-optimal results in supervised learning frameworks [159]. Several studies have also found that most supervised methods lead to a huge performance drop when applied to different centre datasets [159]. Changes in patient population, the appearance of lesions, imaging modalities used, and differences in hardware all affect data variability, pose a bottleneck during training, and adversely affect model performance. We ask if we can leverage already available high-quality public datasets with and without labels to fine-tune these models without compromising algorithmic performance but instead boosting them.

More recently, the availability of foundational models in endoscopy has further accelerated progress in this field [160, 161], enabling more generalisable and transferable solutions across tasks and datasets. These developments have highlighted the potential of self-supervised learning approaches, which aim to learn meaningful representations from unlabelled data. By leveraging intrinsic image properties or structural cues, self-supervised learning mitigates the reliance on large annotated datasets—an especially valuable advantage in medical imaging where expert-labelled data are scarce and expensive to obtain.

Self-supervised learning (SSL) methods learn semantically meaningful features by training a ML method using unlabelled data first. The pre-trained model is then fine-tuned on a training sample with the available labelled samples for each specific downstream task, thus eliminating the requirement of a large amount of labelled data during training, improving generalisation capability for the next downstream task and expansion to other out-of-distribution datasets [162]. In the medical imaging field, SSL has been used extensively for different tasks, including disease classification [154, 163], lesion region detection [155, 164] and segmentation [165, 166]. Since medical imaging data are not abundantly available, pretext learning tasks in SSL can be used to leverage the learnt representation of the training data to benefit the downstream tasks. For example, [167] used the whole training sets for brain MRI, abdominal CT, and fetal ultrasound images for the pretext learning before fine-tuning on successive downstream tasks.

Endoscopy remains the clinical standard for diagnosing and surveying disease in hollow organs. In contrast to data obtained from other imaging modalities, the analysis of endoscopy video is extremely challenging [156] due to various factors such as internal organ deformation, light interaction with tissue at different depths, imaging artefacts such as bubbles, fluid and other floating objects, and a considerable operator dependency. Subtle and fine-grained changes often indicate the onset of disease. Developing robust computer-aided techniques to detect such changes poses a significant challenge. In this work, we propose a novel approach with a comprehensive loss function that improves learning on the pretext task by using

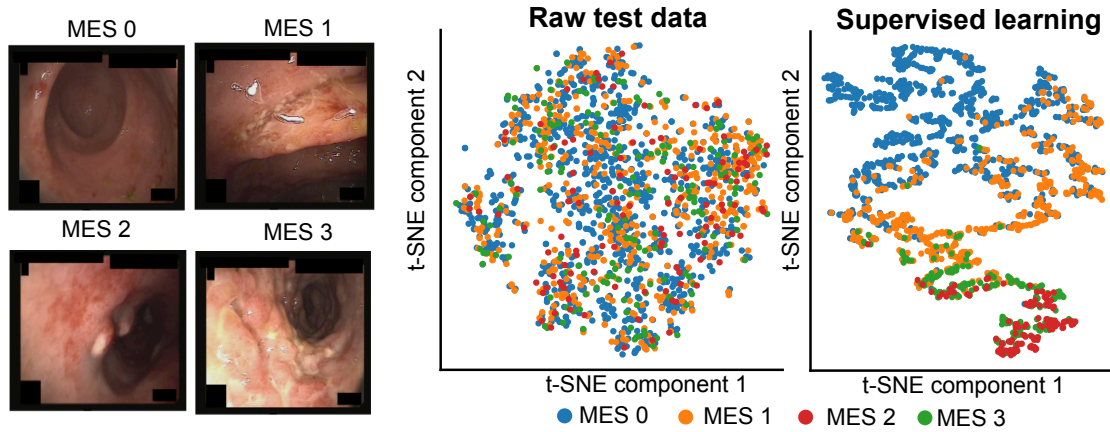


**Figure 4.3:** Endoscopic image analysis for ulcerative colitis scoring. (on left) Representative images for Mayo endoscopic scoring from 0 to 3, and (on right) t-SNE plots [168] with perplexity of 15 and maximum iterations of 1000 for all test samples before learning and after supervised learning using ResNet50 [35].

both image-level and patch-level discrimination. To this extent, we also use memory banks[169] to store positive and negative samples with moving weights that help to learn features that are semantically meaningful for downstream tasks. In our previous work [170], we only explored classification task for downstream task with single disease type and used NCE loss similar to PIRL together with added patch level discrimination only. In this work, we have included classification, detection, and semantic segmentation as downstream tasks and used two different lesion types observed under endoscopic imaging. Further, we introduce a penalisation technique to better exploit inter-class variations in positive and negative samples using additive angular margin in our contrastive loss. By transforming images into jigsaw puzzles and computing contrastive losses between different feature embeddings, we learn a representation capable of differentiating between the subtle characteristics of the different classes. In addition, we also explore the introduction of an attention mechanism[137] in our network for further improvement. Key contributions of our presented work can be summarised below:

- Novel SSL-CPCD method can learn semantically meaningful features from unlabeled data, improving performance on subsequent tasks, including classification, detection, and segmentation of two different lesion types in endoscopic images.

- Single and group-level instances are used to minimise noise contrastive estimation loss, increase inter-class separation, and minimise intra-class distance. For this, we establish a loss between the target image-level and patch-level embedding.
- We propose to include an additive angular margin[171] within the cosine similarity in the contrastive loss to penalise the decision boundary between the positive and the negative samples further, increasing the inter-class separation.
- Evaluation of our method on four different datasets including Kvasir-SEG [172], CVC-ClinicDB [173], LIMUC [147], and our in-house dataset.
- We show that our SSL-CPCD-based method outperforms several SOTA SSL strategies by a large margin.

### 4.2.2 Methodology

We propose a novel self-supervised approach that exploits the invariant representation learning beneficial for downstream tasks by using both image-level instance-group discrimination and patch-level instance-group discrimination losses. To this extent, we propose two novel approaches (see Fig. 4.4) - firstly, exploiting positive and negative samples for noise contrastive loss estimation ( $\mathcal{L}_{\text{NCE}}$ ). Unlike classically used  $\mathcal{L}_{\text{NCE}}$  [174], we integrate an added angular margin [171] computed between the negative samples embedding and learned normalised weights into the NCE loss. This enables the dissociation of different samples further.

Secondly, we employ  $k$ -means clustering and adopt a cross-view approach [128] for both image-level and patch-level instances. This enables group-wise association, indicating that similar embedding belong to distinct groups. A similarity matrix score is then computed for each sample  $i$  between image-level cluster centroid  $C_i^k$  and normalised patch-level feature embedding  $\bar{g}_i$ , i.e.,  $\langle \bar{g}_i, C_i^k \rangle$ , and between patch-level cluster centroid  $PC_i^k$  and normalised image-level feature embedding  $\bar{f}_i$ , i.e.,  $\langle \bar{f}_i, PC_i^k \rangle$ . This determines how close the embedding of the same class  $k$  ( $k$ -means clustered) are in either image instances  $C_i^k$  or patch instances  $PC_i^k$ . Thus,



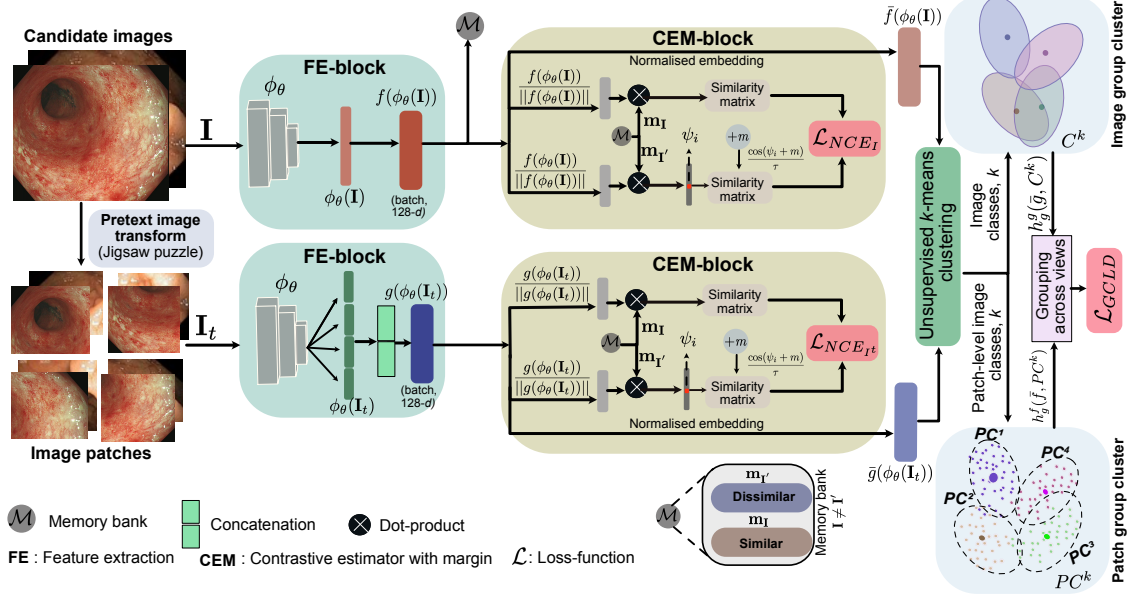


**Figure 4.4:** Block diagram of our proposed self-supervised learning framework with composite pretext-class discrimination losses. The ResNet50 encoder network [35] is fine-tuned using original images with transformations and randomly shuffled patches, simulating a jigsaw puzzle task as in [120], within a self-supervised setting. This setup encourages the model to learn spatially invariant and discriminative features, improving generalisability and performance in downstream tasks. A contrastive estimator with margin is separately computed for both image-level and patch-level instances. Additionally, a group-wise contrastive loss is introduced by comparing the centroids at the patch level ( $PC^k$ ) and the image level ( $C^k$ ). A memory bank  $\mathcal{M}$  [169] is utilised to store all feature representations for efficient contrastive learning.

we apply a cross-view approach for such a similarity association by computing a similarity matrix between the centroid of patch cluster labels with each corresponding cluster label at the image level and vice-versa (detailed below (Section B-E)).

Our novel group-wise loss enables us to learn fine-grained features at both patch-level  $\mathcal{L}_{PC}^k$  and image level  $\mathcal{L}_C^k$  that can enhance more local representations. For grouping of the embeddings, here we utilise a  $k$ -means clustering technique with class numbers similar to downstream tasks to provide representative clusters. Our approach uses memory banks[169] to store all representations useful for various loss function estimations. Below we have described each element of our approach presented in the block diagram in Fig. 4.4.

#### 4.2.2.1 Feature extraction (FE) block

Let the endoscopy dataset  $\mathcal{D}$  consist of  $N$  image samples, denoted as  $\mathcal{D} = \{\mathbf{I}_1, \mathbf{I}_2, \dots, \mathbf{I}_N\}$ . We use a set of image transformations  $\mathcal{T}$  to create and reshuffle  $m$  number of image patches for each image in the dataset ( $\mathcal{D}, \mathcal{P} = \{\mathbf{I}_{1t}^1, \dots, \mathbf{I}_{1t}^m, \dots, \mathbf{I}_{Nt}^1, \dots, \mathbf{I}_{Nt}^m\}$  with  $t \in \mathcal{T}$ , where  $t$  represents a transformation matrix applied to image  $\mathbf{I}$  such that a transformed image  $\mathbf{I}_t$  is obtained). We train a convolutional neural network (*in our case*, ResNet50[35]) with free parameters  $\theta$  that embody representation  $\phi_\theta(\mathbf{I})$  for a given sample  $\mathbf{I}$  and  $\phi_\theta(\mathbf{I}_t)$  for patches  $\mathcal{P}$ .

**Image-level embedding:** Candidate images are fed in batches which are transformed using simple geometric (horizontal and vertical flips) and photometric (colour jitter with 0.4 for hue, saturation, contrast and brightness) transformations and fed into an encoder giving a feature representation  $\phi_\theta(\mathbf{I})$ . We then apply a projection head  $f(\cdot)$  to re-scale the representations to a 128-dimensional feature vector.

**Patch-level embedding:** Each image is divided into nine patches and randomly shuffled to create transformation of patches that mimic shuffled “Jigsaw puzzle” pieces[120]. But unlike solving a jigsaw pretext task [127], here we use these patches as transformed image patches. In this case, we perform random cropping and shuffle of the cropped areas into patch size of  $64 \times 64$  along with the colour transforms used for the original images. The transformation included random horizontal and vertical flips, and photo-metric changes such as brightness, contrast, hue and saturation with a factor of 0.4 (changes between 60% and 140% of the original image). Representations of each patch constituting the image  $\mathbf{I}$  are concatenated to form  $\phi_\theta(\mathbf{I}_t)$ . A projection head  $g(\cdot)$ [121] is applied to re-scale the representations to a 128-dimensional feature vector.

**Memory banks:** The memory bank[169]  $\mathcal{M}$  stores all the feature representations of the dataset  $\mathcal{D}$  at the image level computed from the original images  $\mathbf{I}$ . These embedding weights are moving average of feature representation  $f(\phi_\theta(\mathbf{I}))$  represented as  $\mathbf{m}_\mathbf{I}$  with assigned indexes that helps to build negative samples  $\mathbf{m}_{\mathbf{I}'}$  for each image during contrastive loss estimation.  $\mathcal{M}$  is updated at every epoch with the step-size of  $0.5 \times$  initial weight and normalised to between 0 and 1 similar to [120].

#### 4.2.2.2 Contrastive loss estimation with margin (CEM) block

Noise contrastive estimator (NCE) [120, 174, 175] is used to measure the similarity scores. In our noise contrastive estimator, let the positive sample pair be  $f_i$  and  $f_i^+$  where  $f_i$  be the normalised positive sample representation of an instance and  $f_i^+$  be the target normalised feature embedding from the moving average  $\mathbf{m}_I$ , and  $\mathcal{D}_p \subseteq \mathcal{D}$  where  $D_p$  denote number of positive samples. Similarly, let  $f_i^-$  be the negative sample representations from the moving average representations  $\mathbf{m}_{I'}$  and  $\mathcal{D}_n \subseteq \mathcal{D}$  with  $D_n$  denoting number of negative samples. Then, following prior works[120, 175], the NCE loss models the posterior probability  $h(f_i, f_i^+)$  of sample  $i$  given a data distribution  $\mathcal{D}$  and temperature parameter  $\tau$  as:

$$h(f_i, f_i^+) = \left( \frac{\exp \frac{\langle f_i, f_i^+ \rangle}{\tau}}{\exp \frac{\langle f_i, f_i^+ \rangle}{\tau} + \sum_{f_i^- \in \mathcal{D}_n} \exp \frac{\langle f_i, f_i^- \rangle}{\tau}} \right) \quad (4.2)$$

However, unlike [120, 175], we propose to add an angular margin[171] to increase the separation between the target embedding  $f_i$  and the “negative samples”  $\mathbf{m}_{I'}$  in our contrastive loss estimation with margin block (CEM-block). We do this by first computing angular separation between the positive target embedding and negative embedding ( $\psi$ )  $\psi = \arccos \langle f_i, \mathbf{m}_{I'} \rangle$  and then add an angular margin  $m$  to the computed angle, i.e.  $\psi_{new} = \psi + m$ . Finally, *cosine* of the  $\psi_{new}$  gives our new similarity between the positive and negative samples. The same CE block is applied for both image-level and patch-level NCE loss computations. One main motivation behind adding the angular margin  $m$  is to increase the gap between the positive and negative samples for both image-level and patch-level feature embedding used in discriminative contrastive loss functions. Adding a constant  $m$  pushes the decision boundary between the negative samples and the target sample further enabling the network to learn improved inter class separation.

Thus, Eq. 4.3 represents the posterior probability with added angular margin[171]. Here, the angular additive margin aims to further increase the separation between feature representations of target embedding and negative samples.

$$h^{new}(f_i, f_i^+) = \left( \frac{\exp \frac{\langle f_i, f_i^+ \rangle}{\tau}}{\exp \frac{\langle f_i, f_i^+ \rangle}{\tau} + \sum_{f_i^- \in \mathcal{D}_n} \exp \frac{\cos(\psi_{new})}{\tau}} \right) \quad (4.3)$$

with  $\psi_{new} = \arccos(\langle f_i, f_i^- \rangle) + m$

The total NCE loss entails minimising the joint-loss (logarithmic) function of Eq. 4.3 both at the image-level and patch-level configurations (also see Fig. 4.4) which are our another contribution. Thus, if  $\bar{f}(= \frac{f}{\|f\|})$  and  $\bar{g}(= \frac{g}{\|g\|})$  are the normalised feature embeddings for the target image and target patch then the total NCE loss can be written as:

$$\mathcal{L}_{NCE}^{total}(\mathbf{I}, \mathbf{I}_t) = \lambda \mathcal{L}_{NCE_I}(\mathbf{m}_I, \bar{f}(\phi_\theta(\mathbf{I}))) + (1 - \lambda) \mathcal{L}_{NCE_{I_t}}(\mathbf{m}_I, \bar{g}(\phi_\theta(\mathbf{I}_t))). \quad (4.4)$$

Each NCE loss component can be established as a negative log-posterior distribution of data samples and negative samples[120, 169, 174]. We compute the NCE losses at the image-level  $\mathcal{L}_{NCE_I}$  and at the patch-level  $\mathcal{L}_{NCE_{I_t}}$  (also see Fig. 4.4) using posterior probability in Eq. 4.3 given as below [169]:

$$\mathcal{L}_{NCE_I}(\mathbf{m}_I, \bar{f}(\phi_\theta(\mathbf{I}))) = -\log[h^{new}(\bar{f}(\phi_\theta(\mathbf{I})), \mathbf{m}_I)] - \sum_{\mathbf{I}' \in \mathcal{D}_n} \log[1 - h^{new}(\mathbf{m}_{I'}, \bar{f}(\phi_\theta(\mathbf{I})))], \quad (4.5)$$

$$\& \mathcal{L}_{NCE_{I_t}}(\mathbf{m}_I, \bar{g}(\phi_\theta(\mathbf{I}_t))) = -\log[h^{new}(\bar{g}(\phi_\theta(\mathbf{I}_t)), \mathbf{m}_I)] - \sum_{\mathbf{I}' \in \mathcal{D}_n} \log[1 - h^{new}(\mathbf{m}_{I'}, \bar{g}(\phi_\theta(\mathbf{I}_t)))]. \quad (4.6)$$

Since, we take the logarithmic of function in Eq. 4.3, the denominator term comparing the positive and negative embeddings with added angular margin acts as a regularisation function. The configured joint-loss  $\mathcal{L}_{NCE}^{total}(\mathbf{I}, \mathbf{I}_t)$  enables to learn representations of image  $\mathbf{I}$  closer to its transformed counterpart  $\mathbf{I}_t$  of the same instance and also to the memory representation  $\mathbf{m}_I$  that will damp the parameter updates in the weights  $\phi_\theta$ . It also further penalises the representations at both image and patch-level from other set of negative images  $\mathbf{I}'$ .

#### 4.2.2.3 $k$ -means feature grouping

One important limitation of single instance discrimination as done in NCE loss is that they focus on within-instance similarity by data augmentation assuming a single distinctive instance, but in downstream tasks, these can appear as various similar observations of the same instance. Thus, a grouping strategy can help mitigate such limitations, as presented in this Section.

**Normalised projection head:** We utilise the linear projection heads to normalise the feature embedding with  $l_2$ -norm[121] that enables to reduce variance from data augmentation and maps the features onto a unit hypersphere,  $\bar{f}(\phi_\theta(\mathbf{I})) = \frac{f(\phi_\theta(\mathbf{I}))}{\|f(\phi_\theta(\mathbf{I}))\|}$ , and  $\bar{g}(\phi_\theta(\mathbf{I}_t)) = \frac{g(\phi_\theta(\mathbf{I}_t))}{\|g(\phi_\theta(\mathbf{I}_t))\|}$ .

**Feature grouping:** To overcome the limitation of the single instance approach, we have used grouping instances based on the local clusters within a batch of samples similar to[128]. We create  $k$  clusters where  $k$  is the number of classes (say  $n$ ) in the downstream tasks and use this to define clusters at image and patch levels. Using spherical  $k$ -means clustering, we group the unit-length feature vectors. We compute the cluster centroids for each image embedding  $C^k$  and patch embedding  $PC^k$  in batch input with  $k = \{1, \dots, n\}$ , where  $n$  is the number of cluster classes depending on the downstream task. We assign each instance in the image and patches to each of their corresponding nearest centroids, say  $C(i) = j$ , meaning instance  $i$  is assigned to centroid  $j$  and so on.

#### 4.2.2.4 Cross-level discrimination at image and patch-levels

**Cross-level grouping:** Clusters could be noisy, so we applied a cross-view local group for each instance by an element-wise multiplication of the feature embedding at image-level  $\bar{f}(\phi_\theta(\mathbf{I}))$  with the cluster centroid of image patches  $PC_i^k$ , and at patch-level  $\bar{g}(\phi_\theta(\mathbf{I}_t))$  with the cluster centroid  $C_i^k$  of the images in the batch where  $i$  is the feature embedding assigned to the cluster.

**Cross-level contrastive loss:** The noise contrastive estimation (NCE) loss [174] across the views can be defined using the expression in Eq. (4.2). However, here, we will use the group cluster embeddings and centroids, and we want to assume

that the group in the patch-level cluster is identical to the group in image level for that specific class. Thus, the cross-level grouping of image-level representation compared to patch-level centroid can be defined with the contrastive loss as[128]:

$$h_g^f(\bar{f}_i, PC_i) = -\log \frac{\exp \frac{\langle \bar{f}_i, PC_i \rangle}{\tau}}{\exp \frac{\langle \bar{f}_i, PC_i \rangle}{\tau} + \sum_{j \neq i} \exp \frac{\langle \bar{f}_i, PC_j \rangle}{\tau}} \quad (4.7)$$

Similarly, our cross-level grouping of the patch-level representations can be also written as:

$$h_g^g(\bar{g}_i, C_i) = -\log \frac{\exp \frac{\langle \bar{g}_i, C_i \rangle}{\tau}}{\exp \frac{\langle \bar{g}_i, C_i \rangle}{\tau} + \sum_{j \neq i} \exp \frac{\langle \bar{g}_i, C_j \rangle}{\tau}} \quad (4.8)$$

The final group-wise cross-level discrimination loss  $\mathcal{L}_{GCLD}$  incorporating both image-level and patch-level representations in combined form with weight  $\lambda$  can be written as:

$$\mathcal{L}_{GCLD} = \sum_{k=1}^k \sum_{i=1}^N \{ \lambda h_g^f(\bar{f}_i, PC_i^k) + (1 - \lambda) h_g^g(\bar{g}_i, C_i^k) \} \quad (4.9)$$

Each patch-level and image-level instance are already compared independently within our CEM-block (see Fig. 4.4 and Eq. 4.4) so using group-level association between image and patch clusters can learn to discriminate features that were not established in the CEM-block. Associating images with patch clusters and vice-versa at a group level in Eq. 4.9 can enhance the model's understanding of context and relationships between different parts of an image. The loss function encourages similarity and consistency between associated images and patch clusters. By clustering patch-level embedding and associating them with the image itself, the model is encouraged to learn consistent representations within a local context. This helps create a metric space where the same image-level and patch-level instances get closer, facilitating better generalisation.

#### 4.2.2.5 Proposed loss

Our final novel loss function that defines single instance-level, group-level, and cross-level representations as a joint loss optimisation problem is referred to as composite pretext-class discrimination loss (CPCD). In this work, noise contrastive

**Table 4.3:** Colonoscopic datasets used in our experiments

| Dataset                                  | Images | Input size       | Train | Valid | Test |
|------------------------------------------|--------|------------------|-------|-------|------|
| <b>Ulcerative colitis classification</b> |        |                  |       |       |      |
| LIMUC [147]                              | 11276  | $224 \times 224$ | 8631  | 959   | 1686 |
| In-house                                 | 251    | $224 \times 224$ | 0     | 0     | 251  |

estimation loss refers to the single instance-level representations and the group-wise cross-level discrimination loss. Empirically,  $\lambda$  in Eq. 4.4 and Eq. 4.9 are set to 0.5 to balance the impact of patch-level and image-level embedding. Thus, the final CPCD loss  $\mathcal{L}_{CPCD}$  combines Eq. 4.4 and Eq. 4.9 with  $\lambda'$  as weighting factor and is given as:

$$\mathcal{L}_{CPCD} = \sum_{k=1}^k \lambda' \underbrace{\sum_{i=1}^N 0.5 \cdot \{h_g^f(\bar{f}_i, PC_i^k) + h_g^g(\bar{g}_i, C_i^k)\}}_{\mathcal{L}_{GCLD}} + (1 - \lambda') \underbrace{\sum_{i=1}^N 0.5 \cdot \{\mathcal{L}_{NCE_I}(\mathbf{m}_I, \bar{f}_i(\phi_\theta(\mathbf{I}))) + \mathcal{L}_{NCE_{I^t}}(\mathbf{m}_I, \bar{g}_i(\phi_\theta(\mathbf{I}^t)))\}}_{\mathcal{L}_{NCE}} \quad (4.10)$$

### 4.2.3 Experiments

#### 4.2.3.1 Dataset

We have explored various colonoscopic imaging datasets that are available publicly and in-house for three different downstream tasks. For the classification task, LIMUC [147] and one in-house dataset (collected under universal patient consenting at the Translational Gastroenterology Unit, John Radcliffe Hospital, Oxford) are applied. It is well-established that clinical decision-making is based on visually good quality images during colonoscopy surveillance and that despite the patient variability certain traits for each specific disease, such as ulcerative colitis and/or polyps remain very closely similar. Hence, the quality of data used in this work corresponds to the data acquired in clinical procedures that make up the majority of cases. The details about the datasets and the number of training, validation, and testing samples used are presented in Table 4.3. All datasets are publicly available including the in-house dataset used for generalisability assessment.

#### 4.2.3.2 Evaluation metrics

We have used standard top- $k$  accuracy (percentage of samples predicted correctly, top1 and top2 are used), F1-score ( $= \frac{2tp}{2tp+fp+fn}$ , tp: true positive, fp: false positive), specificity ( $= \frac{tp}{tp+fn}$ ), sensitivity or recall ( $= \frac{tn}{tn+fp}$ ), and Quadratic Weighted Kappa (QWK) for our classification task.

#### 4.2.3.3 Implementation details

The proposed method is implemented using PyTorch [176]. All experiments were conducted on an NVIDIA Quadro RTX 6000 graphics card. For pretext tasks in self-supervised learning, we have used the batch size of 32 and all models in the experiments were trained all models until convergence with the largest number of epochs set to 2000. The SGD optimiser with a learning rate of  $1e^{-3}$  was used for training and was empirically set. All input images were resized to  $224 \times 224$  pixels. ImageNet pre-trained model weights were used during pretext training.

For the downstream classification task, we fine-tuned the model with a learning rate of  $1e^{-4}$ , the SGD optimiser with a batch size of 32, and a stopping criteria with the patience of 20 epochs. For all downstream tasks, our proposed model converged below 200 epochs. The same stopping criteria as in the fine-tuning approach was applied for the baseline methods.

The training of baseline networks included different data augmentation techniques such as geometric augmentation including random rotation ( $\pm 5^\circ$ ), random horizontal flip (probability,  $p = 0.5$ ), random crop of input size with padding adding of 10 pixels, random affine transformation ( $10^\circ$ ) and photometric augmentation including brightness, contrast, and saturation changes between 50% and 150% of the original image. The same code was used for both fine-tuning of SSL approaches and baseline.

All experiments used 80% of the dataset for training, 10% for validation, and the remaining held-out 10% for testing. We additionally have used out-of-centre unseen centre datasets for generalisability study. We also provide experiments on 10%, 20%



**Table 4.4:** Quantitative comparison for UC classification task on LIMUC dataset

| Method                | Backbone | Top 1         | Top 2         | F1            | Spec.         | Recall        | QWK           | P-values  |
|-----------------------|----------|---------------|---------------|---------------|---------------|---------------|---------------|-----------|
| Baseline [35]         | R50      | 0.7532        | 0.9346        | 0.6689        | 0.8630        | 0.7011        | 0.8237        | 7.022e-07 |
| Baseline [35, 137]    | R50-Att. | 0.7556        | 0.9414        | 0.6727        | 0.8692        | 0.7029        | 0.8290        | 4.342e-06 |
| SimCLR [121]          | R50      | 0.7355        | 0.9387        | 0.6631        | 0.8510        | 0.6752        | 0.8083        | 6.527e-11 |
| SimCLR [121]          | R50-Att. | 0.7384        | 0.9219        | 0.6649        | 0.8431        | 0.6942        | 0.8102        | 4.843e-09 |
| SimCLR+DCL [177]      | R50      | 0.7555        | 0.9450        | 0.6729        | 0.8635        | 0.6897        | 0.8269        | 1.506e-07 |
| SimCLR+DCL [177]      | R50-Att. | 0.7568        | 0.9367        | 0.6755        | 0.8669        | 0.6952        | 0.8287        | 8.760e-07 |
| MoCoV2+CLD [128]      | R50      | 0.7574        | 0.9493        | 0.6788        | 0.8721        | 0.6959        | 0.8309        | 7.691e-12 |
| MoCoV2+CLD [128]      | R50-Att. | 0.7598        | 0.9536        | 0.6812        | 0.8709        | 0.7047        | 0.8333        | 6.787e-09 |
| PIRL [120]            | R50      | 0.7651        | 0.9637        | 0.6859        | 0.8874        | 0.7098        | 0.8376        | 4.201e-04 |
| PIRL [120]            | R50-Att. | 0.7740        | 0.9610        | 0.6918        | 0.8893        | 0.7133        | 0.8460        | 1.621e-03 |
| PIRL+PLD (ours) [170] | R50      | 0.7752        | 0.9666        | 0.7040        | 0.8891        | 0.7129        | 0.8509        | 6.043e-03 |
| PIRL+PLD (ours) [170] | R50-Att. | 0.7847        | 0.9707        | 0.7167        | 0.8933        | 0.7146        | 0.8563        | 2.967e-02 |
| SSL-CPCD (ours)       | R50      | 0.7912        | 0.9633        | 0.7209        | <b>0.9043</b> | 0.7198        | 0.8693        | -         |
| SSL-CPCD (ours)       | R50-Att. | <b>0.7977</b> | <b>0.9750</b> | <b>0.7279</b> | 0.9008        | <b>0.7259</b> | <b>0.8746</b> | -         |

and 50% training data for fine-tuning. To guarantee experimental reproducibility we have conducted all experiments by setting random seed at 42.

**Hyperparameters:** For group-wise cross-level discrimination loss ( $\mathcal{L}_{GCLD}$  in Eq. (4.9)), we set  $k = 4$  for a number of clusters in classification pretext task,  $k = 2$  in detection and segmentation pretext task,  $s = 6$  for the re-scaling and  $m = 0.5$  for an angular margin. Memory bank settings proposed in [120] has been used with the same hyperparameters. For Eq. (4.4) we use  $\lambda = 0.5$  and use  $\tau = 0.4$  for computing the function  $h(.,.)$  in Eq. (4.2, 4.3, 4.7, and 4.8). We used an updated weight of 0.5 for the memory bank exponential moving average representations. These values are justified in our ablation study provided in Section 4.2.4.4. Hyperparameter tuning for our SSL parameters ( $\tau$ , and  $\lambda$ ) is conducted on the downstream tasks utilising the validation set which consists of labeled samples. Further, we have performed hyperparameter tuning (such as learning rate) with the same validation set for each method to report the best performance of each on this dataset.

#### 4.2.4 Results

In this section, we present the comparison of our proposed SSL-CPCD approach with other SOTA SSL methods.

#### 4.2.4.1 Comparison for UC classification task

ResNet50 [35] (R50) and ResNet50 with convolution-block attention module[137] (R50-Att.) are established as the baseline model for supervised learning first, and then the same is used for other SOTA SSL-based method comparisons in Table 4.4 for ulcerative colitis classification task on LIMUC dataset. Baseline networks R50 and R50-Att., respectively, obtained 75.39% and 75.62% on top-1 accuracy and 82.51% and 82.78% on QWK. Our proposed SSL-CPCD method yielded the best results with 79.77%, 72.79%, 90.08%, 72.59% and 87.46% on top 1 accuracy, F1 score, specificity, recall and QWK, respectively. Compared to the supervised learning-based baseline models (R50), the top 1 accuracy and QWK is improved by 4.38% and 4.95%, respectively, using our proposed SSL-CPCD with the same backbones. We also compared our proposed SSL-CPCD approach with other SOTA SSL methods, including popular SimCLR [121], SimCLR+DCL [177], MoCoV2+CLD [122] and PIRL [120] methods. Our proposed network (R50-Att.) outperformed all these methods with at least nearly 2.4% (PIRL) up to 6% (SimCLR) on top-1 accuracy. Similar improvements can be observed on other metrics as well.

#### 4.2.4.2 Computation time

For the PIRL approach (top SOTA approach) it took 37 s per epoch for pretext training, while ours (SSL-CPCD) took 40 s per epoch with a total of nearly 22.5 hrs. However, for the fine-tuning of the downstream tasks, all methods reported in this chapter with the same backbone took the same amount of time per epoch compared to baseline and SOTA models (e.g., for classification 31s per epoch, detection took 27s per epoch and segmentation approach took 29 s per epoch). With our model, the fine-tuning approach converged at 200 epochs compared to others (e.g., PIRL took 500 epochs for convergence).

#### 4.2.4.3 Generalisation

To ensure the generalisation of the proposed approach, we trained our model and other methods on one dataset and then tested them on an unseen dataset from

**Table 4.5:** Generalisation study for the UC classification task

| Method                | Backbone | Top 1         | Spec.         | Recall        | QWK           | P-values  |
|-----------------------|----------|---------------|---------------|---------------|---------------|-----------|
| Baseline [35]         | R50      | 0.5856        | 0.7239        | 0.5569        | 0.5379        | 2.315e-11 |
| Baseline [35, 137]    | R50-Att. | 0.6055        | 0.7539        | 0.5739        | 0.6572        | 6.812e-08 |
| SimCLR[121]           | R50      | 0.5737        | 0.7020        | 0.5256        | 0.5611        | 6.076e-09 |
| SimCLR[121]           | R50-Att. | 0.5777        | 0.7139        | 0.5420        | 0.6018        | 1.497e-06 |
| SimCLR+DCL[177]       | R50      | 0.5976        | 0.7297        | 0.5622        | 0.6345        | 4.277e-07 |
| SimCLR+DCL[177]       | R50-Att. | 0.6016        | 0.7458        | 0.5758        | 0.6542        | 9.603e-06 |
| MoCoV2+CLD[128]       | R50      | 0.6175        | 0.7716        | 0.5878        | 0.6939        | 5.657e-08 |
| MoCoV2+CLD[128]       | R50-Att. | 0.6135        | 0.7823        | 0.5737        | 0.6902        | 6.069e-11 |
| PIRL[120]             | R50      | 0.6255        | 0.8213        | 0.6097        | 0.7312        | 1.370e-03 |
| PIRL[120]             | R50-Att. | 0.6335        | 0.8397        | 0.6139        | 0.7469        | 9.677e-04 |
| PIRL+PLD (ours) [170] | R50      | 0.6370        | 0.8359        | 0.6197        | 0.7507        | 2.574e-02 |
| PIRL+PLD (ours) [170] | R50-Att. | 0.6453        | 0.8415        | 0.6223        | 0.7662        | 5.047e-03 |
| SSL-CPCD(ours)        | R50      | 0.6534        | 0.8501        | 0.6249        | 0.7835        | -         |
| SSL-CPCD(ours)        | R50-Att. | <b>0.6733</b> | <b>0.8677</b> | <b>0.6403</b> | <b>0.7887</b> | -         |

different institutions. Generalisability study for UC classification We used the UC classification model trained on the LIMUC dataset collected at Marmara University School of Medicine. We tested this model on our in-house dataset (collected at the John Radcliffe Hospital, Oxford). Table 4.5 assesses the generalisability of our SSL-CPCD model and other SOTA approaches on UC classification task. Our proposed SSL-CPCD obtained an acceptable Top 1 accuracy of 67.33%, F1-score of 64.69%, specificity of 86.77%, recall of 64.03% and QWK of 78.87%. While outperforming all SOTA approaches, compared with MoCoV2+CLD, our method achieves an improvement of 5.98% on top 1 accuracy and nearly 9% in QWK. Table 4.5 shows that our SSL-CPCD outperforms other SOTA methods in various evaluation metrics.

#### 4.2.4.4 Ablation studies

We have conducted an extensive ablation study of our approach. First, we ablated the impact of multiple loss functions, including NCE, GCLD, and the added angular margin  $m$ . Then, we conducted an ablation study experiment to further evaluate the performance of our proposed approach under different parameter settings.

**Loss functions:** Table 4.6 shows the quantitative results of our ablation study in loss functions on classification tasks. Initially, our proposed method, which contains

**Table 4.6:** Ablation study results for different loss functions on validation set

| Loss function | Class. task   |               |
|---------------|---------------|---------------|
|               | Top 1         | F1            |
| NCE           | 0.8341        | 0.7693        |
| NCE+GCLD      | 0.8495        | 0.781         |
| NCE+GCLD+ $m$ | <b>0.8603</b> | <b>0.7933</b> |

**Table 4.7:** Effect of different hyper-parameter setting using validation set

| Parameter settings |        | Top 1         |
|--------------------|--------|---------------|
| $\lambda'$         | $\tau$ |               |
| 0.1                | 0.2    | 0.8342        |
| 0.25               | 0.2    | 0.8415        |
| 0.5                | 0.2    | 0.8509        |
| 1                  | 0.2    | 0.8405        |
| 0.1                | 0.4    | 0.8425        |
| 0.25               | 0.4    | 0.8467        |
| 0.5                | 0.4    | <b>0.8603</b> |
| 1                  | 0.4    | 0.8446        |
| 0.1                | 0.6    | 0.8332        |
| 0.25               | 0.6    | 0.8383        |
| 0.5                | 0.6    | 0.8498        |
| 1                  | 0.6    | 0.8352        |

**Table 4.8:** Hyper-parameter study for different number of clusters

| No. of clusters | Top 1         |
|-----------------|---------------|
| $k$             |               |
| 2               | 0.8533        |
| 3               | 0.8540        |
| 4               | <b>0.8603</b> |
| 5               | 0.8582        |
| 7               | 0.8498        |
| 10              | 0.8519        |

three loss functions, achieves 79.12% on top 1 accuracy and 72.09% on the F1 score for the classification task. It can be observed that compared with classically using noise contrastive loss only, our approach and modifications led to significant improvements in all downstream tasks by a larger margin (top 1 accuracy and F1 score improved respectively by 2.61% and 2.4%).

**Impact of hyper-parameters:** The quantitative results for the ablation study of different parameter settings are shown in Table 4.7. We set different weight

**Table 4.9:** Abalation study for different learning rate using validation set

| Method             | LR           | Top 1         | F1            | Spec.         | Recall        | QWK           |
|--------------------|--------------|---------------|---------------|---------------|---------------|---------------|
| SimCLR [121]       | 0.05         | 0.7987        | 0.7298        | 0.9214        | 0.7340        | 0.8707        |
|                    | 0.01         | 0.8010        | 0.7317        | 0.9220        | 0.7379        | 0.8732        |
|                    | 0.005        | 0.8021        | 0.7326        | 0.9246        | <b>0.7433</b> | 0.8751        |
|                    | <b>0.001</b> | <b>0.8049</b> | <b>0.7345</b> | <b>0.9239</b> | 0.7414        | <b>0.8769</b> |
|                    | 0.0005       | 0.8004        | 0.7301        | 0.9206        | 0.7355        | 0.8716        |
|                    | 0.0001       | 0.7955        | 0.7270        | 0.9189        | 0.7328        | 0.8647        |
| SSL-CPCD<br>(ours) | 0.05         | 0.8501        | 0.7817        | 0.9469        | 0.7869        | 0.9253        |
|                    | 0.01         | 0.8559        | 0.7899        | 0.9502        | 0.7873        | 0.9299        |
|                    | 0.005        | 0.8567        | 0.7905        | 0.9507        | 0.7889        | 0.9305        |
|                    | <b>0.001</b> | <b>0.8603</b> | <b>0.7933</b> | <b>0.9523</b> | <b>0.7911</b> | <b>0.9326</b> |
|                    | 0.0005       | 0.8533        | 0.7882        | 0.9487        | 0.7891        | 0.9279        |
|                    | 0.0001       | 0.8429        | 0.7846        | 0.9453        | 0.7865        | 0.9231        |

**Table 4.10:** Comparision with SOTA methods for different percentages of training samples for fine-tuning in classification as a downstream task. ResNet50 is taken as baseline for all cases.

| Method                      | % training samples | Top 1         | F1            | Spec.         | Recall        | QWK           |
|-----------------------------|--------------------|---------------|---------------|---------------|---------------|---------------|
| Baseline<br>(ResNet50) [35] | 100                | 0.7539        | 0.6702        | 0.8670        | 0.6906        | 0.8251        |
|                             | 50                 | 0.7462        | 0.6580        | 0.8602        | 0.6732        | 0.8187        |
|                             | 20                 | 0.7263        | 0.6432        | 0.8513        | 0.6655        | 0.8065        |
|                             | 10                 | 0.7004        | 0.6277        | 0.8242        | 0.6387        | 0.7879        |
| PIRL [120]                  | 100                | 0.7651        | 0.6859        | 0.8874        | 0.7098        | 0.8376        |
|                             | 50                 | 0.7236        | 0.6472        | 0.8324        | 0.6588        | 0.7993        |
|                             | 20                 | 0.6903        | 0.6363        | 0.8196        | 0.6350        | 0.7832        |
|                             | 10                 | 0.6813        | 0.6139        | 0.8017        | 0.6247        | 0.7649        |
| PIRL+PLD<br>[170]           | 100                | 0.7752        | 0.7040        | 0.8891        | 0.7129        | 0.8509        |
|                             | 50                 | 0.7497        | 0.6719        | 0.8560        | 0.6861        | 0.8196        |
|                             | 20                 | 0.7301        | 0.6636        | 0.8455        | 0.6727        | 0.8043        |
|                             | 10                 | 0.7135        | 0.6359        | 0.8397        | 0.6518        | 0.7937        |
| SSL-CPCD<br>(ours)          | 100                | <b>0.7912</b> | <b>0.7209</b> | <b>0.9043</b> | <b>0.7198</b> | <b>0.8693</b> |
|                             | 50                 | <b>0.7811</b> | <b>0.7133</b> | <b>0.8977</b> | <b>0.7158</b> | <b>0.8562</b> |
|                             | 20                 | <b>0.7692</b> | <b>0.6978</b> | <b>0.8749</b> | <b>0.6897</b> | <b>0.8242</b> |
|                             | 10                 | <b>0.7551</b> | <b>0.6897</b> | <b>0.8666</b> | <b>0.6709</b> | <b>0.8001</b> |

and temperature in Eq. (4.4-4.9). Weight parameter  $\lambda' = \{0.1, 0.25, 0.5, 1\}$  and temperature  $\tau = \{0.2, 0.4, 0.6\}$  are used for searching best parameters experimentally. As shown in Table 4.7, when weight and temperature parameters are 0.5 and 0.4, respectively, our method achieves the best results in classification tasks with 79.77% on Top 1 accuracy. We additionally explored different learning rates for all compared

**Table 4.11:** Comparison with larger baseline backbones for classification, detection, and segmentation tasks.

| Method                             | Parameters | Top 1  |
|------------------------------------|------------|--------|
| EfficientNet-v2 [178] (Baseline)   | 24M        | 0.7562 |
| EfficientNet-v2 [178] (+ SSL-CPCD) | 24M        | 0.7948 |
| EfficientNet-B6 [179] (Baseline)   | 43M        | 0.7556 |
| EfficientNet-B6 [179] (+ SSL-CPCD) | 43M        | 0.7953 |
| ResNet101 [35] (Baseline)          | 44.5M      | 0.7550 |
| ResNet101 [35] (+ SSL-CPCD)        | 44.5M      | 0.7930 |
| ResNet152 [35] (Baseline)          | 60.2M      | 0.7568 |
| ResNet152 [35] (+ SSL-CPCD)        | 60.2M      | 0.7972 |
| SENet-154 [180] (Baseline)         | 440M       | 0.7612 |
| SENet-154 [180] (+ SSL-CPCD)       | 440M       | 0.8021 |

methods, and our experiments were conducted on a validation set that suggested the same learning rate of 0.001 (Table 4.9). To substantiate the effect of the number of clusters on cross-level grouping for the loss computation, we conducted an additional study (Table 4.8). It can be observed that the best results were obtained when using the same number of clusters as the number of classes in downstream tasks.

**Limited data settings:** It can be observed that compared to the baseline method and most accurate SSL technique (PIRL [120] and PIRL+PLD [170]), our approach outperforms in all data settings for all cases. For example, in classification task (Table 4.10) with only 10% data our SSL-CPCD approach outperforms the supervised ResNet50 trained with 100% data, while only 50% data was required to outperform other two most accurate SSL methods.

**Comparison with other larger models:** It can be observed from Table 4.11 that our approach (SSL-CPCD) consistently provides a performance boost of nearly 4% in classification independent of model size.

**Model efficiency:** Compared to the baseline SSL PIRL model [120], the proposed SSL-CPCD demonstrated faster convergence both during the pretext learning stage and the fine-tuning stage (e.g., loss at 100th epoch from the proposed approach of 2.16 vs 3.36 during pretext, and 0.11 vs 0.35 on the fine-tuning stage). While the inference time from all methods was the same for all SSL-based methods as the baseline supervised methods, our SSL-CPCD took nearly 22.2 hours (compared to 20 hours 30 minutes with the baseline PIRL) for pretext training and nearly 1 hour

40 minutes for fine-tuning for all downstream tasks (31s per epoch for classification, 27s per epoch for detection and 29s per epoch for segmentation tasks).

#### 4.2.4.5 Qualitative Analysis

In the UC classification task in Fig. 4.5, a t-distributed stochastic neighbour embedding (*t*-SNE) plot of test image samples embedding, and gradient weighted activation map (Grad-CAM) method is used to visualise model performance. It can be observed in Fig. 4.5 a) that more compact and interpretable clusters were obtained using perplexity of 15 compared to perplexity of 5. It can be observed (Fig. 4.5 a.ii) that the test images are stochastically distributed in raw sample distributions. It is important to note that data points close to each other in the *t*-SNE visualisation are likely to be similar in the original higher dimensional embedding space[168]. This means that points forming tight clusters are more similar than points in other clusters. It can be observed that after model training, images of the same class cluster appear tentatively in the same region. The SSL-CPCD method that utilises group-wise loss based on clustering demonstrates improved grouping and their distinct separation from other clusters in different MES classes than the baseline supervised model and SSL-based PIRL approach. Using SSL-CPCD, it can be observed that the same categories are more concentrated in the same area, and there are clear boundaries between different categories, which in other cases are not apparent.

This can also be quantified with the cluster distances between the classes presented for ResNet50 (supervised), PIRL (SSL), and our approach. Our SSL-CPCD method improves the inter-class separation, demonstrating the ability of our approach to discriminate different class representations more clearly. For example, compared to the most accurate SOTA SSL-PIRL the cluster distance between MES 0 and MES 1, MES 1 and MES2, and MES 2 and MES 3 are increased by 10.95%, 15.01%, and 52.11%, respectively. As observed in Fig.4.5 a), MES 1 is also confused with MES 3 in some cases. For this case, our approach increases the cluster distance by 18.24%.

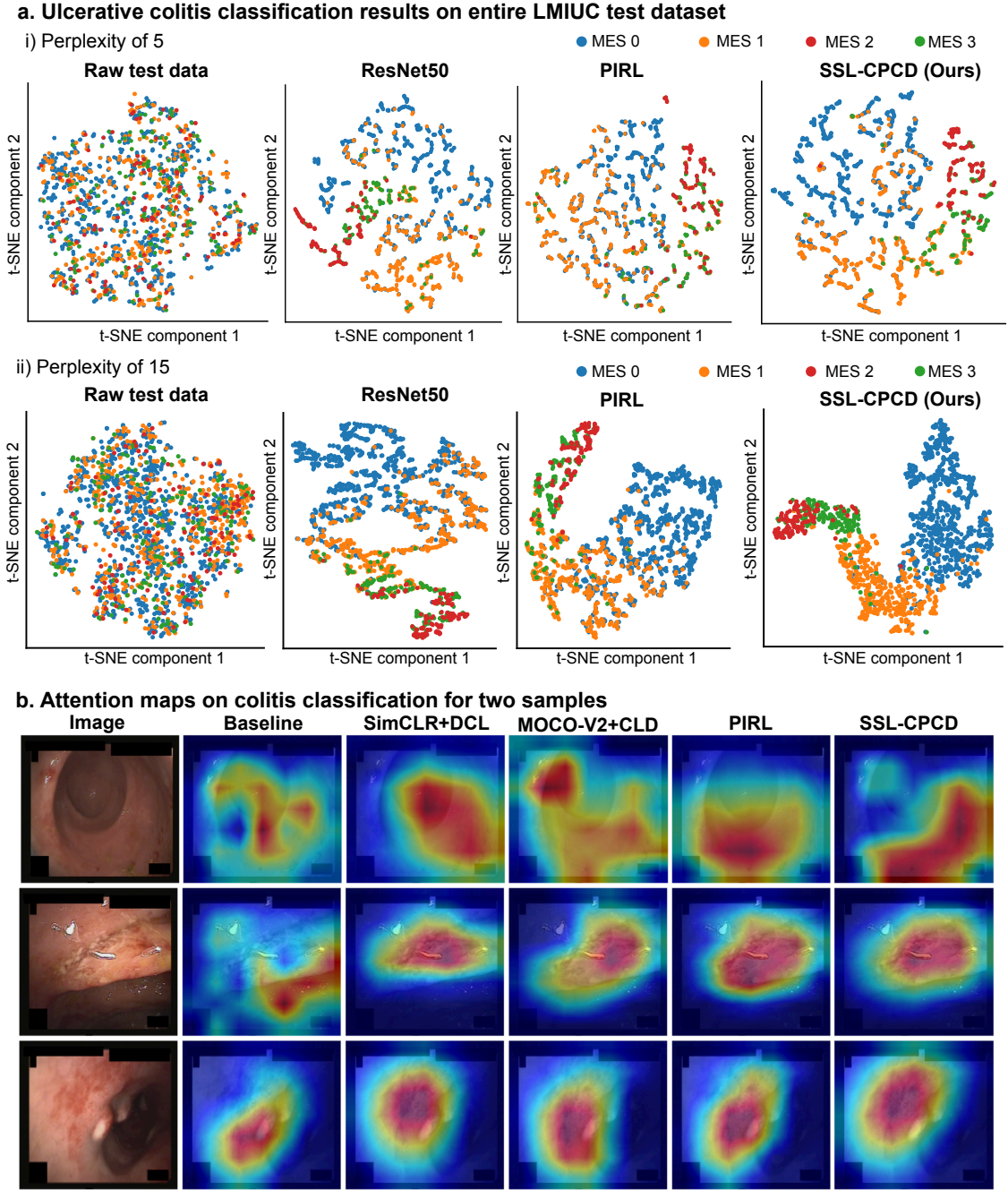


**Figure 4.5:** a) t-SNE plot for the raw test data, baseline network (supervised) and two SSL approaches with perplexity of 5 in a. i) and 15 in a. ii); and b) attention maps of the proposed approach compared to other state-of-the-art methods for multi-class ulcerative colitis classification task (MES 0, MES 1, and MES 2). Principal Components Analysis was used to reduce the dimension of the feature maps to 50 as suggested in[168]. It is to be noted that t-SNE plots were obtained before 1000 iterations (set maximum) as the t-SNE converged, i.e., the same error after a certain number of iterations. The minimisation cost (Kullback-Leibler divergence) error was higher for all plots for the perplexity of 5 compared to the perplexity of 15.



Similarly, while looking at the attention (Fig. 4.5 b), the baseline method focuses on the wrong location in some images (see the first and second rows in Fig. 4.5 b). In other SOTA SSL methods, the model notices the correct location, but the lesion location is inaccurate. Our proposed SSL-CPCD can accurately identify the severely affected lesion area and shape.

#### 4.2.5 Contribution and conclusion

While supervised learning methods have been widely used in the endoscopic image analysis, however, due to limited labelled data availability and large variability in disease-relevant changes in the tissue structure or used imaging devices at different centres, their generalisability can be largely affected [181, 182].

We explored a self-supervised-based learning approach (SSL) that can learn semantically meaningful features and representations invariant to texture and illumination changes in endoscopic images that are more robust. SSL can learn robust features as they are not supervised or incentivised by labels under the pretext of training. Thus, it can learn label-relevant features during fine-tuning and other intrinsic properties that may help generalise better [157]. In addition, self-supervised approaches are a scalable way to learn visual representations without labels and are well-known to be robust to class imbalance compared to supervised approaches [157]. We use patch generation during pretext training first that is independent to the class labels in the dataset (i.e., unlabelled samples are used). Here, patches will have their own labels, e.g., solving generated as a jigsaw puzzle in our case, and the model tries to learn representations by learning the similarity and dissimilarity between the patches and the target image. So, there is no notion of class imbalance within pretext training. The learned visual representations during pretext tasks help to fine-tune the model performance on the downstream task more effectively and are widely known to be robust to class imbalance in the dataset as suggested by Liu *et al.* [157]. Our work is validated on inconspicuous ulcerative colitis data and more visible polyps. This gives us an insight into the strength of the performance of SSL approaches on variable data and target lesion types.

We show that these representations, using unlabeled endoscopic images, mitigate the risk of limited labels and provide improved results compared to widely used supervised techniques. Even though the SSL-based approaches have been proposed in the past for natural scenes [120–122], to our knowledge, no study has been conducted comprehensively for endoscopic image analysis. We propose a novel composite pretext-class discrimination loss (CPCD) that combines noise contrastive losses for the single instance level and group-based instance, showing significant improvements compared to other SSL methods. Here, instance discrimination obtains meaningful representations through instance-level contrastive learning, which can be used to reflect the apparent similarities between instances.

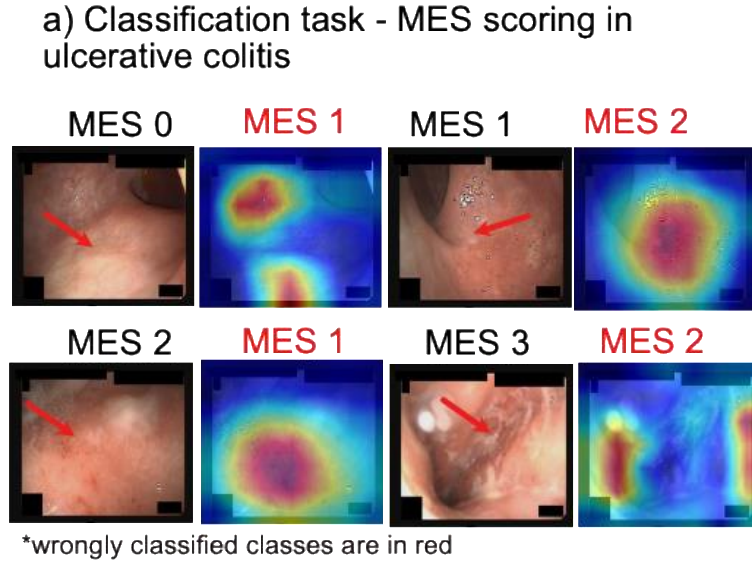


**Figure 4.6:** Limitations of our proposed approach for different data conditions, including blur, difficult scene views due to organ topology or camera positions, imaging artefacts such as specularities and pixel saturation.

The assumption that instance discrimination is established is based on the fact that each example is significantly different from others and can be treated as a separate category. However, endoscopic image data tend to have higher similarity in their video images, making it extremely hard to learn reliable features. Thus, there is a significant similarity between training data in conventional self-supervised learning, which will lead to the negative pairs used in the contrastive learning

process being likely to be composed of high similarity instances, which will lead to a large number of false positives in the training process of contrastive learning repulsion. We solve this problem in two directions. First, we propose a patch-level instance-group discrimination, GCLD loss, which can perform  $k$ -means clustering on instances so that similar instances are clustered into the same group. The error rejection of high-similarity instances was alleviated in the subsequent contrastive loss. In addition, we further optimise the loss function by adding an angular margin  $m$  between positive and negative samples in contrastive learning (see ablation study results in Table 4.6). Our proposed SSL-CPCD significantly improves representative tasks for anomalies in colonoscopy images. In the ulcerative colitis classification task, SSL-CPCD succeeded with the highest Top 1 accuracy of 79.77% and the highest F1 score of 72.79% on LIMUC (see Table 4.4). Furthermore, SSL-CPCD on the generalisability assessment it achieves the highest Top 1 accuracy and QWK of 67.33% and 78.87% (see Table 4.5).

Additionally, P-values from paired t-tests show statistical significance in results from our methods compared to others. It can be observed that for UC classification tasks our approach provided P-value  $\leq 0.05$  (significantly different, Table 4.4).

However, due to organ topology and the complex environment of moving the endoscopic camera, there are unavoidable artefacts, blur due to camera motion and differences in visual appearances that the training samples may not sufficiently capture. We investigated which of the frames gave lower scores in each downstream task (see Figure 4.6). It can be observed that for the classification task in Figure 4.6 (a) inaccurate results are mainly in the frames where the organ topology is complex (e.g., lifted mucosa from surrounding in the first case and the fourth case). Similarly, the blur in sample 3, labelled MES 2, is identified as MES 1.

We used various techniques to tackle the implementation challenges of SSL approaches. For example, in terms of the requirement of large amounts of data to generate pseudo labels, we used larger datasets that are publicly available and then used testing with a held-out smaller dataset. We also applied various percentage splits on it during fine-tuning training (e.g., 50%, 20% and 10%) to measure the

effectiveness of the method and its ability to exploit learnt representations. We observed a similar trend of performance drop to baseline upon reducing label samples from 100% to 10%, which was also observed in [154]. However, it is essential to note that our SSL-CPCD approach for classification with only 10% data samples (Table 4.10) can get the performance of fully supervised learning approaches that require 100% labelled data. Similarly, to understand the effect of cluster size  $k$  (Table 4.8) and the various other hyperparameters (Table 4.7), we used a heuristic approach on all SSL approaches for downstream task, and the best values were identified and reported in the chapter. A marginal drop in the performance with different cluster sizes  $k$  was observed compared to the target classes in the downstream task. This can also be viewed as a limitation of the proposed method.

Our proposed approach combining image-level and group-level instances in a contrastive loss-based framework for self-supervised learning in endoscopic image analysis is unique and has not been explored before. Our SSL-CPCD approach can learn representative features from unlabeled images that are evident to improve any downstream tasks. Our strategy of the added angular margin increases the geometric distance between positive and negative samples. Our experiments demonstrate the effectiveness and improvement of our SSL-CPCD method over several SOTA self-supervised methods on three downstream tasks for complex colonoscopic images. Cross-dataset testing confirmed the generalisation ability of our SSL-CPCD approach, which is superior to all SOTA SSL-based methods.

# 5

## Computer-aided colorectal polyp segmentation and detection

### Contents

---

|            |                                                                                                                                       |            |
|------------|---------------------------------------------------------------------------------------------------------------------------------------|------------|
| <b>5.1</b> | <b>Motivation . . . . .</b>                                                                                                           | <b>96</b>  |
| <b>5.2</b> | <b>Results from baseline and proposed self-supervised approach . . . . .</b>                                                          | <b>97</b>  |
| 5.2.1      | Experiments . . . . .                                                                                                                 | 97         |
| 5.2.1.1    | Baseline and Comparison Methods . . . . .                                                                                             | 97         |
| 5.2.1.2    | Dataset . . . . .                                                                                                                     | 97         |
| 5.2.1.3    | Evaluation metrics . . . . .                                                                                                          | 98         |
| 5.2.1.4    | Implementation details . . . . .                                                                                                      | 98         |
| 5.2.2      | Results . . . . .                                                                                                                     | 99         |
| 5.2.2.1    | Comparison for polyp detection task . . . . .                                                                                         | 99         |
| 5.2.2.2    | Comparison for polyp segmentation task . . . . .                                                                                      | 99         |
| 5.2.2.3    | Generalisability study for polyp segmentation . . . . .                                                                               | 100        |
| 5.2.2.4    | Ablation studies . . . . .                                                                                                            | 101        |
| 5.2.2.5    | Qualitative Analysis . . . . .                                                                                                        | 103        |
| 5.2.3      | Contribution and conclusion . . . . .                                                                                                 | 104        |
| <b>5.3</b> | <b>End-to-end self-supervised and temporal self-attention with feature branching for real-time video polyp segmentation . . . . .</b> | <b>108</b> |
| 5.3.1      | Introduction . . . . .                                                                                                                | 108        |
| 5.3.2      | Methodology . . . . .                                                                                                                 | 110        |
| 5.3.2.1    | Encoder self-supervision . . . . .                                                                                                    | 111        |
| 5.3.2.2    | Normalised self-attention block . . . . .                                                                                             | 112        |
| 5.3.2.3    | Global-to-local feature learning . . . . .                                                                                            | 113        |
| 5.3.2.4    | Decoder . . . . .                                                                                                                     | 113        |
| 5.3.2.5    | Joint-loss optimisation (end-to-end framework) . . . . .                                                                              | 114        |

|         |                                          |     |
|---------|------------------------------------------|-----|
| 5.3.3   | Experimental setup . . . . .             | 114 |
| 5.3.3.1 | Datasets and Comparison models . . . . . | 114 |
| 5.3.3.2 | Evaluation metrics . . . . .             | 115 |
| 5.3.3.3 | Implementation details . . . . .         | 115 |
| 5.3.4   | Results . . . . .                        | 116 |
| 5.3.4.1 | Method comparisons . . . . .             | 116 |
| 5.3.4.2 | Generalization . . . . .                 | 118 |
| 5.3.4.3 | Ablation Study . . . . .                 | 120 |
| 5.3.4.4 | Qualitative results . . . . .            | 122 |
| 5.3.5   | Discussion . . . . .                     | 122 |
| 5.3.6   | Conclusion . . . . .                     | 125 |

---

This chapter presents polyp detection and segmentation task based on deep learning. In Section 5.2, the results of some baseline models and applying the methods in Section 4.2 to downstream polyp detection and segmentation tasks are shown. Subsequently, we introduced a novel real-time video polyp segmentation method that is robust and accurate over existing state-of-the-art methods in Section 5.3.

Originality: For more comprehensive experiments, we used three public datasets including Kvasir-SEG [172], CVC-ClinicDB [173] and SUN-SEG [183]. In Section 5.3, We explore a novel architecture with self-supervised learning as an auxiliary task and a spatial-temporal self-attention mechanism for improved representation learning integrated within the network. Our novel end-to-end configuration and joint optimisation of both self-supervised contrastive loss and cross-entropy losses enable the network to learn more discriminative contextual features in the videos. This work is under review by Neurocomputing.

- **Xu, Ziang**, Jens Rittscher, and Sharib Ali. "SSTFB: Leveraging self-supervised pretext learning and temporal self-attention with feature branching for real-time video polyp segmentation." arXiv preprint arXiv:2406.10200, 2024.(Neurocomputing-in review)

## 5.1 Motivation

The study described in Section 5.2 repurposes the approach introduced in Chapter 4 as a preliminary investigation into polyp segmentation and detection tasks. While

this method demonstrated strong performance in static polyp segmentation, its application to video-based polyp segmentation revealed notable limitations—particularly in maintaining accuracy across temporally adjacent frames. This performance gap highlighted the challenges of extending static segmentation models to dynamic, real-world endoscopic video sequences. These observations motivated the development of a more specialised framework tailored for video polyp segmentation, as presented in Section 5.3. The proposed approach addresses the identified shortcomings by incorporating temporal modelling techniques to improve consistency and robustness across frames, filling a gap in existing literature that often overlooks video-level segmentation.

## 5.2 Results from baseline and proposed self-supervised approach

### 5.2.1 Experiments

#### 5.2.1.1 Baseline and Comparison Methods

For both the polyp detection and segmentation tasks, we employ RetinaNet [65] and Res-UNet [184] as the respective baseline models. In each case, we evaluate two backbone configurations: ResNet50 [35] and ResNet50 integrated with a Convolutional Block Attention Module (CBAM) [137]. All models are pretrained using the same self-supervised learning strategy described in Section 4.2, and subsequently fine-tuned on their respective downstream tasks. Detailed fine-tuning configurations are provided in Section 5.2.1.4.

Consistent with the experiments conducted in Section 4.2, we also perform extensive evaluations using other SOTA SSL methods as comparison baselines. These include the widely adopted SimCLR [121], SimCLR+DCL [177], MoCoV2+CLD [128] and PIRL [120].

#### 5.2.1.2 Dataset

Kvasir-SEG [172] and CVC-ClinicDB [173] are used for the segmentation task. Similarly, Kvasir-SEG [172] for experiments on polyp detection as a downstream

**Table 5.1:** Colonoscopic datasets used in our experiments

| Dataset                   | Images | Input size       | Train | Valid | Test |
|---------------------------|--------|------------------|-------|-------|------|
| <b>Polyp segmentation</b> |        |                  |       |       |      |
| Kvasir-SEG [172]          | 1000   | Variable         | 800   | 100   | 100  |
| SUN(non-polyp) [183]      | 1000   | Variable         | 1000  | 0     | 0    |
| CVC-ClinicDB [173]        | 612    | $384 \times 288$ | 0     | 0     | 612  |
| <b>Polyp detection</b>    |        |                  |       |       |      |
| Kvasir-SEG [172]          | 1000   | Variable         | 800   | 100   | 100  |
| SUN [183]                 | 1000   | Variable         | 1000  | 0     | 0    |

task. In the pretext task for the detection and segmentation tasks, we have used polyp samples from Kvasir-SEG [172] and 1000 non-polyp samples from the SUN dataset [183] for training our SSL model. This approach ensures the presence of representative images for both polyp and non-polyp samples during pretext task learning, maintaining a balanced representation. Non-polyp frames were randomly sampled from 10 video IDs (starting from ID 1), with a selection of 100 frames per video, incorporating variability in non-polyp categories across the available 109,554 frames. We used a random seed of 42, which guarantees the reproducibility of this selection process. The details about the datasets and the number of training, validation, and testing samples used are presented in Table 5.1.

#### 5.2.1.3 Evaluation metrics

For the detection task, standard computer vision metrics, including mean average precision (mAP at an IoU interval [0.25:0.05:0.75]) and AP small, medium and large, were used for our experiments. Dice similarity coefficient (DSC), which is also known as F1-score, and type-II error referred to as F2-score, recall and positive predictive values ( $PPV, = \frac{tp}{tp+fp}$ ) have been used for evaluating our segmentation task.

#### 5.2.1.4 Implementation details

We take the same setting for pretext tasks in Section 4.2. For the detection task, we have used the Adam optimiser with a learning rate of  $1e^{-5}$  and a batch size of 32 with a learning rate decay of 0.1 with patience 3 and 400 epochs. For the segmentation task, 300 epochs with a batch size of 16 and an SGD optimiser



with initial learning rate of  $1e^{-3}$  and a learning rate decay of 0.9 times per 20 epochs were used to fine-tune the model.

### 5.2.2 Results

**Table 5.2:** Quantitative comparison for polyp detection task using Kvasir-SEG dataset

| Method                | Backbone | mAP                                | AP25          | AP50          | AP75          | APsmall       | APmedium      | APlarge       | P-values  |
|-----------------------|----------|------------------------------------|---------------|---------------|---------------|---------------|---------------|---------------|-----------|
| RetinaNet [65]        | R50      | 0.8637 $\pm$ 0.101                 | 0.9377        | 0.8965        | 0.6973        | 0.4832        | 0.7507        | 0.8398        | 6.150e-03 |
| RetinaNet [65]        | R50-Att. | 0.8729 $\pm$ 0.082                 | 0.9436        | 0.9097        | 0.7045        | 0.4871        | 0.7603        | 0.8419        | 1.263e-03 |
| SimCLR [121]          | R50      | 0.8501 $\pm$ 0.090                 | 0.9259        | 0.8837        | 0.6709        | 0.4641        | 0.7416        | 0.8237        | 1.337e-05 |
| SimCLR [121]          | R50-Att. | 0.8532 $\pm$ 0.076                 | 0.9278        | 0.8818        | 0.6846        | 0.4679        | 0.7403        | 0.8302        | 9.270e-03 |
| SimCLR+DCL [177]      | R50      | 0.8537 $\pm$ 0.087                 | 0.9269        | 0.8853        | 0.6852        | 0.4709        | 0.7429        | 0.8321        | 5.057e-04 |
| SimCLR+DCL [177]      | R50-Att. | 0.8592 $\pm$ 0.101                 | 0.929         | 0.8893        | 0.6887        | 0.4739        | 0.7467        | 0.8403        | 2.156e-04 |
| MoCoV2+CLD [128]      | R50      | 0.8457 $\pm$ 0.072                 | 0.9273        | 0.9028        | 0.6845        | 0.4779        | 0.7491        | 0.8346        | 1.869e-09 |
| MoCoV2+CLD [128]      | R50-Att. | 0.8519 $\pm$ 0.070                 | 0.9412        | 0.9044        | 0.7019        | 0.4859        | 0.7563        | 0.8402        | 7.792e-07 |
| PIRL [120]            | R50      | 0.8612 $\pm$ 0.078                 | 0.9403        | 0.8931        | 0.6929        | 0.4839        | 0.7487        | 0.8317        | 2.707e-02 |
| PIRL [120]            | R50-Att. | 0.8677 $\pm$ 0.073                 | 0.9408        | 0.8961        | 0.702         | 0.4863        | 0.7589        | 0.8408        | 8.771e-03 |
| PIRL+PLD (ours) [170] | R50      | 0.8649 $\pm$ 0.077                 | 0.9431        | 0.9027        | 0.7042        | 0.4919        | 0.7547        | 0.8426        | 5.694e-03 |
| PIRL+PLD (ours) [170] | R50-Att. | 0.8738 $\pm$ 0.071                 | 0.9466        | 0.9053        | 0.7069        | 0.4977        | 0.7601        | 0.8458        | 1.178e-02 |
| SSL-CPCD (ours)       | R50      | 0.8709 $\pm$ 0.070                 | 0.9421        | 0.9192        | 0.7107        | 0.5033        | 0.763         | 0.8542        | -         |
| SSL-CPCD (ours)       | R50-Att. | <b>0.8862<math>\pm</math>0.067</b> | <b>0.9469</b> | <b>0.9227</b> | <b>0.7197</b> | <b>0.5105</b> | <b>0.7703</b> | <b>0.8598</b> | -         |

#### 5.2.2.1 Comparison for polyp detection task

The Kvasir-SEG polyp dataset was used to evaluate the performances of SSL on detection as the downstream task in endoscopy. Here, we have chosen RetinaNet [65] as the baseline network for both the supervised and the self-supervised learning approaches. The quantitative results from Table 5.2 show that our proposed SSL-CPCD approach outperforms all the other SOTA methods on all metrics. It achieves 2.29%, 2.7% and 3.3% improvement on mAP compared to SSL methods, including MoCoV2+CLD, SimCLR+DCL and SimCLR, respectively. Our method also improves 1.83% on AP50 and 1.4% on APmedium (medium polyp sizes) compared to MoCoV2+CLD, respectively. Compared to the widely used supervised technique RetinaNet, our method is better on mean average mAP but significantly improves over AP50, AP75 and size-based metrics.

#### 5.2.2.2 Comparison for polyp segmentation task

The Kvasir-SEG dataset was also used to assess the performance of SSL-based approaches in our experiment for segmentation as a downstream task in endoscopy. Here, we have chosen Res-UNet[184] as the baseline network for both the supervised

**Table 5.3:** Quantitative comparison for polyp segmentation task

| Method                | Backbone | DSC                                | F2            | Recall        | PPV           | P-values  |
|-----------------------|----------|------------------------------------|---------------|---------------|---------------|-----------|
| U-Net [96]            | none     | 0.7933 $\pm$ 0.053                 | 0.7671        | 0.7945        | 0.9131        | 2.881e-10 |
| Res-UNet [184]        | R50      | 0.7867 $\pm$ 0.074                 | 0.7667        | 0.7723        | 0.9139        | 1.670e-04 |
| Res-UNet [184]        | R50-Att. | 0.792 $\pm$ 0.066                  | 0.7743        | 0.7862        | 0.9187        | 7.117e-06 |
| SimCLR [121]          | R50      | 0.7892 $\pm$ 0.068                 | 0.7621        | 0.7639        | 0.9146        | 2.936e-07 |
| SimCLR [121]          | R50-Att. | 0.7945 $\pm$ 0.058                 | 0.7759        | 0.7903        | 0.9162        | 6.977e-05 |
| SimCLR+DCL [177]      | R50      | 0.7879 $\pm$ 0.056                 | 0.7609        | 0.7653        | 0.9169        | 4.690e-05 |
| SimCLR+DCL [177]      | R50-Att. | 0.7933 $\pm$ 0.051                 | 0.7822        | 0.7741        | 0.9038        | 8.816e-06 |
| MoCoV2+CLD [128]      | R50      | 0.7946 $\pm$ 0.093                 | 0.779         | 0.7846        | 0.9173        | 3.697e-11 |
| MoCoV2+CLD [128]      | R50-Att. | 0.8029 $\pm$ 0.077                 | 0.7953        | 0.7998        | 0.9201        | 7.327e-09 |
| PIRL [120]            | R50      | 0.7906 $\pm$ 0.069                 | 0.7842        | 0.7946        | 0.9135        | 7.502e-03 |
| PIRL [120]            | R50-Att. | 0.7969 $\pm$ 0.067                 | 0.7893        | 0.8056        | 0.9177        | 1.432e-02 |
| PIRL+PLD (ours) [170] | R50      | 0.8069 $\pm$ 0.058                 | 0.7913        | 0.8093        | 0.9189        | 5.519e-03 |
| PIRL+PLD (ours) [170] | R50-Att. | 0.8116 $\pm$ 0.049                 | 0.7989        | 0.8132        | 0.9211        | 4.107e-02 |
| SSL-CPCD (ours)       | R50      | 0.8173 $\pm$ 0.047                 | 0.8032        | 0.8104        | 0.9217        | -         |
| SSL-CPCD (ours)       | R50-Att. | <b>0.8232<math>\pm</math>0.043</b> | <b>0.8081</b> | <b>0.8234</b> | <b>0.9259</b> | -         |

and the self-supervised learning approaches. Results for supervised U-Net[96] approach have also been provided for comparison. Table 5.3 compares the result of the proposed SSL-CPCD with other SOTA SSL approaches and baseline supervised model. While proposed SSL-CPCD provided an improvement of 3.12% and 3.72% on DSC and Recall, respectively, for the baseline ResNetUNet in a supervised setting, our approach also showed improvements of 2.03%, 1.28%, 2.36% and 0.58% over MoCoV2 + CLD in DSC, F2-score, recall and PPV, respectively. Higher recall while keeping the precision (PPV) high (over 90%) indicates that our method is more medically relevant.

### 5.2.2.3 Generalisability study for polyp segmentation

All models for both baseline and SOTA approaches were first trained on the Kvasir-SEG dataset and then tested on the CVC-ClinicDB dataset, for which the results are presented in Table 5.4. Our proposed SSL-CPCD drastically surpassed baseline supervised approaches (over 10% on DSC for U-Net and over 7% on DSC with the same backbone on ResUNet). In addition, our method obtained an improvement of 4.73% and 6.8%, respectively, over MoCoV2+CLD and SimCLR on DSC. Similarly, over 5% improvement on PIRL is evident in both backbone settings (R50 and R50-Att.).

**Table 5.4:** Generalisation study for segmentation task

| Method                | Backbone | DSC                 | F2-score      | Recall        | PPV           | P-values  |
|-----------------------|----------|---------------------|---------------|---------------|---------------|-----------|
| U-Net [96]            | none     | 0.5826±0.057        | 0.6029        | 0.5942        | 0.7633        | 1.009-e12 |
| Res-UNet [184]        | R50      | 0.6092±0.040        | 0.6379        | 0.6265        | 0.8218        | 8.708-e05 |
| Res-UNet [184]        | R50-Att. | 0.6027±0.072        | 0.6499        | 0.6372        | 0.8065        | 6.363-e04 |
| SimCLR [121]          | R50      | 0.5942±0.069        | 0.6498        | 0.6334        | 0.8312        | 1.982-e07 |
| SimCLR [121]          | R50-Att. | 0.6113±0.055        | 0.6556        | 0.6673        | 0.8329        | 6.621-e04 |
| SimCLR+DCL[177]       | R50      | 0.6039±0.059        | 0.6501        | 0.6586        | 0.8293        | 8.193-e06 |
| SimCLR+DCL[177]       | R50-Att. | 0.6092±0.049        | 0.6679        | 0.6598        | 0.8301        | 2.770-e03 |
| MOCov2+CLD[128]       | R50      | 0.6268±0.047        | 0.6498        | 0.6509        | 0.8277        | 9.913-e09 |
| MOCov2+CLD[128]       | R50-Att. | 0.632±0.050         | 0.6691        | 0.6675        | 0.8362        | 7.076-e07 |
| PIRL[120]             | R50      | 0.6196±0.049        | 0.6742        | 0.6703        | 0.8378        | 5.501-e03 |
| PIRL [120]            | R50-Att. | 0.6277±0.053        | 0.6801        | 0.6770        | 0.8396        | 6.950-e04 |
| PIRL+PLD (ours) [170] | R50      | 0.6459±0.062        | 0.6617        | 0.6802        | 0.8311        | 1.669e-03 |
| PIRL+PLD (ours) [170] | R50-Att. | 0.6573±0.066        | 0.6720        | 0.6827        | 0.8368        | 9.618e-03 |
| SSL-CPCD (ours)       | R50      | 0.6705±0.043        | 0.6903        | 0.6812        | 0.8379        | -         |
| SSL-CPCD (ours)       | R50-Att. | <b>0.6793±0.040</b> | <b>0.6978</b> | <b>0.6897</b> | <b>0.8488</b> | -         |

#### 5.2.2.4 Ablation studies

**Loss functions:** Table 5.5 shows the quantitative results of our ablation study in loss functions on polyps detection and segmentation tasks. Similarly, it has the best AP50 and mAP of 91.92% and 87.09%, respectively. On the segmentation task, the combined loss also showed improvement when combined with various strategies, yielding 81.13% on DSC and 92.18% on PPV.

**Impact of hyper-parameters:** Same as Section 4.2, We set different weight  $\lambda' = \{0.1, 0.25, 0.5, 1\}$  and temperature  $\tau = \{0.2, 0.4, 0.6\}$  in Eq. (4.4-4.9). As shown in Table 5.6, when weight and temperature parameters are 0.5 and 0.4, respectively, our method achieves the best results in segmentation tasks with 82.32% on DSC. For the detection task, the best performance of our SSL-CPCD was obtained when  $\lambda' = 0.5$  and  $\tau = 0.6$ . We explored the effect of the number of clusters on cross-level grouping loss in Table 5.7,  $k = 2$  got best performances on both

**Table 5.5:** Ablation study results for different loss functions on validation set

| Loss function | Det. task     |               | Seg. task     |               |
|---------------|---------------|---------------|---------------|---------------|
|               | AP50          | mAP           | DSC           | PPV           |
| NCE           | 0.9527        | 0.9004        | 0.8883        | 0.9438        |
| NCE+GCLD      | 0.9613        | 0.9062        | 0.8987        | 0.9569        |
| NCE+GCLD+m    | <b>0.9697</b> | <b>0.9119</b> | <b>0.9037</b> | <b>0.9604</b> |

**Table 5.6:** Effect of different hyper-parameter setting using validation set

| Parameter settings |        | AP50          | DSC           |
|--------------------|--------|---------------|---------------|
| $\lambda'$         | $\tau$ |               |               |
| 0.1                | 0.2    | 0.9461        | 0.8897        |
| 0.25               | 0.2    | 0.9489        | 0.8917        |
| 0.5                | 0.2    | 0.9611        | 0.8968        |
| 1                  | 0.2    | 0.9522        | 0.8903        |
| 0.1                | 0.4    | 0.9503        | 0.8940        |
| 0.25               | 0.4    | 0.9517        | 0.8987        |
| 0.5                | 0.4    | 0.9653        | <b>0.9037</b> |
| 1                  | 0.4    | 0.9549        | 0.8962        |
| 0.1                | 0.6    | 0.9566        | 0.8863        |
| 0.25               | 0.6    | 0.9607        | 0.8907        |
| 0.5                | 0.6    | <b>0.9697</b> | 0.8935        |
| 1                  | 0.6    | 0.9625        | 0.8886        |

**Table 5.7:** Hyper-parameter study for different number of clusters

| No. of clusters | AP50          | DSC           |
|-----------------|---------------|---------------|
| $k$             |               |               |
| 2               | <b>0.9697</b> | <b>0.9037</b> |
| 3               | 0.9578        | 0.9015        |
| 4               | 0.9618        | 0.8999        |
| 5               | 0.9601        | 0.8957        |
| 7               | 0.9627        | 0.9004        |
| 10              | 0.9609        | 0.8986        |

segmentation and detection tasks.

**Limited data settings:** For detection task (Table 5.8) under all data settings, our method outperforms all other approaches with 2%-3% improvements. Finally, for the segmentation downstream task (Table 5.9) where two fully supervised models are used residual U-Net (Res-UNet[184]) and polyp segmentation with pyramid vision transformers (Polyp-PVT[106]), our method surpassed Res-UNet with only 50% data (nearly 2%) and Polyp-PVT trained on 50% data was achieved by only 20% using our approach. Also, it is evident that the supervised method (Res-UNet) dropped DSC score from 0.78 (on 100%) to 0.69 (on 10%), a drop of 9%, while our approach only dropped by 5%. A similar drop was observed for Polyp-PVT (7% compared to only 5%).

**Table 5.8:** Comparison with SOTA methods for different percentages of training samples for fine-tuning in detection as a downstream task. ResNet50 is taken as baseline for all cases.

| Method          | % training samples | mAP                                | AP25          | AP50          | AP75          | P-values  |
|-----------------|--------------------|------------------------------------|---------------|---------------|---------------|-----------|
| RetinaNet [65]  | 100                | 0.8637 $\pm$ 0.101                 | 0.9377        | 0.8965        | 0.6973        | 6.150e-03 |
|                 | 50                 | 0.8321 $\pm$ 0.123                 | 0.9051        | 0.8709        | 0.6779        | 1.479e-05 |
|                 | 20                 | 0.8207 $\pm$ 0.095                 | 0.8778        | 0.8587        | 0.6583        | 3.290e-04 |
|                 | 10                 | 0.8093 $\pm$ 0.117                 | 0.8599        | 0.8327        | 0.6467        | 3.157e-06 |
| PIRL [120]      | 100                | 0.8612 $\pm$ 0.078                 | 0.9403        | 0.8931        | 0.6929        | 2.707e-02 |
|                 | 50                 | 0.8455 $\pm$ 0.081                 | 0.9194        | 0.8793        | 0.6799        | 9.670e-05 |
|                 | 20                 | 0.8260 $\pm$ 0.087                 | 0.8817        | 0.8571        | 0.6642        | 4.551e-04 |
|                 | 10                 | 0.8135 $\pm$ 0.082                 | 0.8627        | 0.8369        | 0.6498        | 7.619e-05 |
| PIRL+PLD [170]  | 100                | 0.8649 $\pm$ 0.077                 | 0.9431        | 0.9027        | 0.7042        | 5.694e-03 |
|                 | 50                 | 0.8398 $\pm$ 0.069                 | 0.9156        | 0.8769        | 0.6826        | 1.191e-03 |
|                 | 20                 | 0.8261 $\pm$ 0.072                 | 0.8942        | 0.8551        | 0.6717        | 2.233e-02 |
|                 | 10                 | 0.8172 $\pm$ 0.079                 | 0.8653        | 0.8390        | 0.6521        | 8.582e-03 |
| SSL-CPCD (ours) | 100                | <b>0.8709<math>\pm</math>0.070</b> | <b>0.9421</b> | <b>0.9192</b> | <b>0.7107</b> | -         |
|                 | 50                 | <b>0.8587<math>\pm</math>0.059</b> | <b>0.9229</b> | <b>0.8917</b> | <b>0.7027</b> | -         |
|                 | 20                 | <b>0.8439<math>\pm</math>0.075</b> | <b>0.9060</b> | <b>0.8764</b> | <b>0.6793</b> | -         |
|                 | 10                 | <b>0.8307<math>\pm</math>0.077</b> | <b>0.8966</b> | <b>0.8671</b> | <b>0.6762</b> | -         |

**Comparison with other larger models:** Compared with the baseline results, our method improved  $\approx 2\%$  in detection and  $\approx 3\%$  in segmentation. (Table 5.10)

### 5.2.2.5 Qualitative Analysis

For the polyp detection task (Fig. 5.1), it can be seen that baseline and other SSL methods cannot accurately locate the polyp’s spatial location. Most methods have enlarged boundaries and even multiple bounding boxes, especially for the second and fourth examples in the figure. However, our SSL-CPCD approach can locate the polyp position more accurately, and the bounding boxes are closer to ground truth.

In the polyp segmentation task (Fig. 5.2), the baseline method incorrectly identifies non-polyp regions as polyps and over or under-segments the area. Although other SSL methods did not misidentify the polyp region, they only segmented part of the polyp. SSL-CPCD can segment polyps more accurately, similar to ground truth labels. Our proposed SSL-CPCD maintains the best segmentation results in all examples.

**Table 5.9:** Comparison with SOTA methods for different percentages of training samples for fine-tuning in segmentation as a downstream task. ResNet50 is taken as a baseline for all cases.

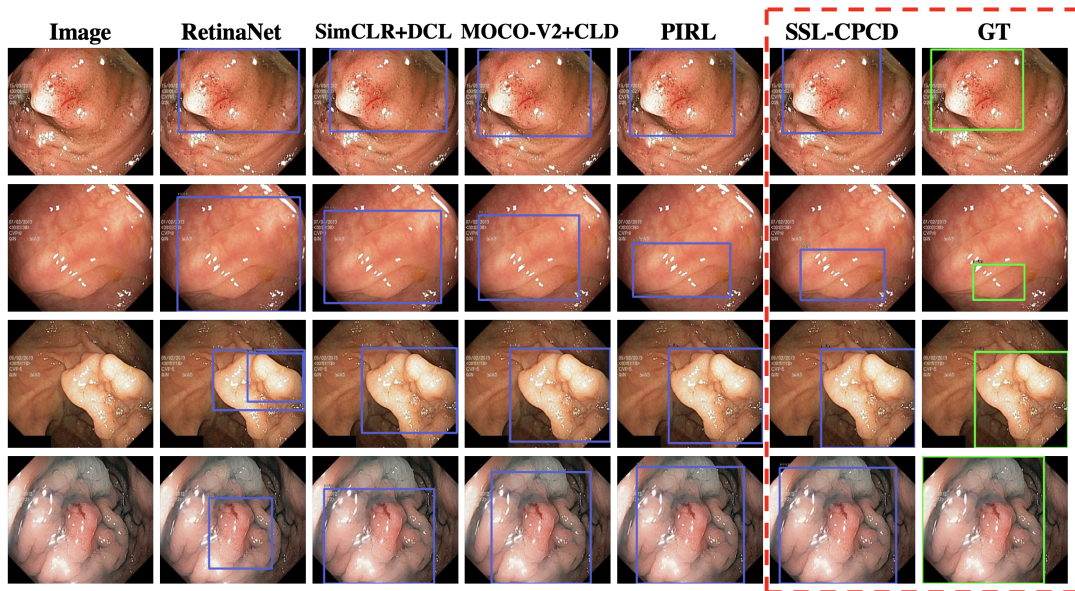
| Method                                    | % training samples | DSC                                | F2            | Recall        | PPV           | P-values  |
|-------------------------------------------|--------------------|------------------------------------|---------------|---------------|---------------|-----------|
| Res-UNet [184]<br>(Supervised)            | 100                | 0.7867 $\pm$ 0.074                 | 0.7667        | 0.7723        | 0.9139        | 1.670e-04 |
|                                           | 50                 | 0.7423 $\pm$ 0.063                 | 0.7341        | 0.7369        | 0.7579        | 2.511e-03 |
|                                           | 20                 | 0.7167 $\pm$ 0.077                 | 0.7062        | 0.7098        | 0.6809        | 9.621e-06 |
|                                           | 10                 | 0.6949 $\pm$ 0.080                 | 0.6897        | 0.6907        | 0.6706        | 5.825e-06 |
| Polyp-PVT [106]<br>(Supervised)           | 100                | 0.9067 $\pm$ 0.042                 | 0.8832        | 0.8897        | 0.9663        | 3.238e-02 |
|                                           | 50                 | 0.8621 $\pm$ 0.053                 | 0.8487        | 0.8479        | 0.9409        | 1.181e-02 |
|                                           | 20                 | 0.8539 $\pm$ 0.059                 | 0.8376        | 0.8301        | 0.9241        | 6.489e-03 |
|                                           | 10                 | 0.8346 $\pm$ 0.057                 | 0.8192        | 0.8221        | 0.9103        | 4.670e-05 |
| PIRL [120]<br>with Res-UNet<br>(SSL)      | 100                | 0.7906 $\pm$ 0.069                 | 0.7842        | 0.7946        | 0.9135        | 7.502e-03 |
|                                           | 50                 | 0.7589 $\pm$ 0.072                 | 0.7433        | 0.7476        | 0.8772        | 9.017e-05 |
|                                           | 20                 | 0.7418 $\pm$ 0.078                 | 0.7306        | 0.7411        | 0.8591        | 1.109e-06 |
|                                           | 10                 | 0.7273 $\pm$ 0.086                 | 0.7139        | 0.7186        | 0.8267        | 8.940e-08 |
| PIRL [120]<br>with Polyp-PVT<br>(SSL)     | 100                | 0.8997 $\pm$ 0.051                 | 0.8781        | 0.8792        | 0.9580        | 2.209e-03 |
|                                           | 50                 | 0.8660 $\pm$ 0.045                 | 0.8465        | 0.8518        | 0.9433        | 4.091e-05 |
|                                           | 20                 | 0.8541 $\pm$ 0.059                 | 0.8329        | 0.8396        | 0.9387        | 4.672e-04 |
|                                           | 10                 | 0.8302 $\pm$ 0.060                 | 0.8201        | 0.8254        | 0.9312        | 2.063e-07 |
| PIRL+PLD [170]<br>with Res-UNet<br>(SSL)  | 100                | 0.8069 $\pm$ 0.058                 | 0.7913        | 0.8093        | 0.9189        | 5.519e-03 |
|                                           | 50                 | 0.7749 $\pm$ 0.053                 | 0.7623        | 0.7646        | 0.8855        | 3.431e-07 |
|                                           | 20                 | 0.7495 $\pm$ 0.067                 | 0.7541        | 0.7372        | 0.8661        | 7.073e-06 |
|                                           | 10                 | 0.7371 $\pm$ 0.069                 | 0.7230        | 0.7287        | 0.8415        | 2.901e-05 |
| PIRL+PLD [170]<br>with Polyp-PVT<br>(SSL) | 100                | 0.9083 $\pm$ 0.049                 | 0.8862        | 0.9007        | 0.9624        | 1.720e-03 |
|                                           | 50                 | 0.8691 $\pm$ 0.052                 | 0.8513        | 0.8577        | 0.9475        | 2.098e-02 |
|                                           | 20                 | 0.8571 $\pm$ 0.055                 | 0.8381        | 0.8431        | 0.9406        | 7.574e-03 |
|                                           | 10                 | 0.8399 $\pm$ 0.063                 | 0.8279        | 0.8316        | 0.9344        | 6.605e-05 |
| SSL-CPCD<br>with Res-UNet<br>(ours)       | 100                | 0.8173 $\pm$ 0.047                 | 0.8032        | 0.8104        | 0.9218        | -         |
|                                           | 50                 | 0.8063 $\pm$ 0.044                 | 0.7867        | 0.7903        | 0.8867        | -         |
|                                           | 20                 | 0.7749 $\pm$ 0.050                 | 0.7647        | 0.7727        | 0.8803        | -         |
|                                           | 10                 | 0.7657 $\pm$ 0.058                 | 0.7562        | 0.7597        | 0.8747        | -         |
| SSL-CPCD<br>with Polyp-PVT<br>(ours)      | 100                | <b>0.9131<math>\pm</math>0.039</b> | <b>0.8960</b> | <b>0.9041</b> | <b>0.9693</b> | -         |
|                                           | 50                 | <b>0.8793<math>\pm</math>0.037</b> | <b>0.8702</b> | <b>0.8833</b> | <b>0.9567</b> | -         |
|                                           | 20                 | <b>0.8697<math>\pm</math>0.040</b> | <b>0.8499</b> | <b>0.8669</b> | <b>0.9508</b> | -         |
|                                           | 10                 | <b>0.8542<math>\pm</math>0.047</b> | <b>0.8407</b> | <b>0.8486</b> | <b>0.9387</b> | -         |

### 5.2.3 Contribution and conclusion

Likewise, we reported the highest values of 88.62%, 94.69%, and 92.27% for mAP, AP25, and AP50 on Kvasir-SEG in the polyp detection task (see Table 5.2). Furthermore, we report the best DSC, recall and PPV for the polyp segmentation task on the Kvasir-SEG dataset (see Table Table 5.3). Furthermore, SSL-CPCD on the generalisability assessment achieving the highest DSC of 67.93% (see Table 5.4).

**Table 5.10:** Comparison with larger baseline backbones for classification, detection, and segmentation tasks.

| Method                             | Parameters | AP50   | DSC    |
|------------------------------------|------------|--------|--------|
| EfficientNet-v2 [178] (Baseline)   | 24M        | 0.9090 | 0.7957 |
| EfficientNet-v2 [178] (+ SSL-CPCD) | 24M        | 0.9221 | 0.8277 |
| EfficientNet-B6 [179] (Baseline)   | 43M        | 0.9028 | 0.7911 |
| EfficientNet-B6 [179] (+ SSL-CPCD) | 43M        | 0.9207 | 0.8203 |
| ResNet101 [35] (Baseline)          | 44.5M      | 0.8993 | 0.7890 |
| ResNet101 [35] (+ SSL-CPCD)        | 44.5M      | 0.9198 | 0.8149 |
| ResNet152 [35] (Baseline)          | 60.2M      | 0.9047 | 0.7938 |
| ResNet152 [35] (+ SSL-CPCD)        | 60.2M      | 0.9214 | 0.8229 |
| SENet-154 [180] (Baseline)         | 440M       | 0.9138 | 0.8016 |
| SENet-154 [180] (+ SSL-CPCD)       | 440M       | 0.9343 | 0.8353 |

**Figure 5.1:** Qualitative comparison of our proposed self-supervised approach with other state-of-the-art methods for polyp detection tasks.

From Fig. 5.2, it is clear that the boundary margins are improved using SSL methods compared to fully supervised methods. This demonstrates that the SSL leverages unlabeled data during pretext learning, encouraging the model to capture semantically meaningful information and robust features from data that is then transferred to the downstream image segmentation task, improving the precision of segmentation boundaries.

P-values from paired t-tests in Table 5.2 and Table 5.3 show significantly different with other methods. It can also be observed that the standard deviation provided

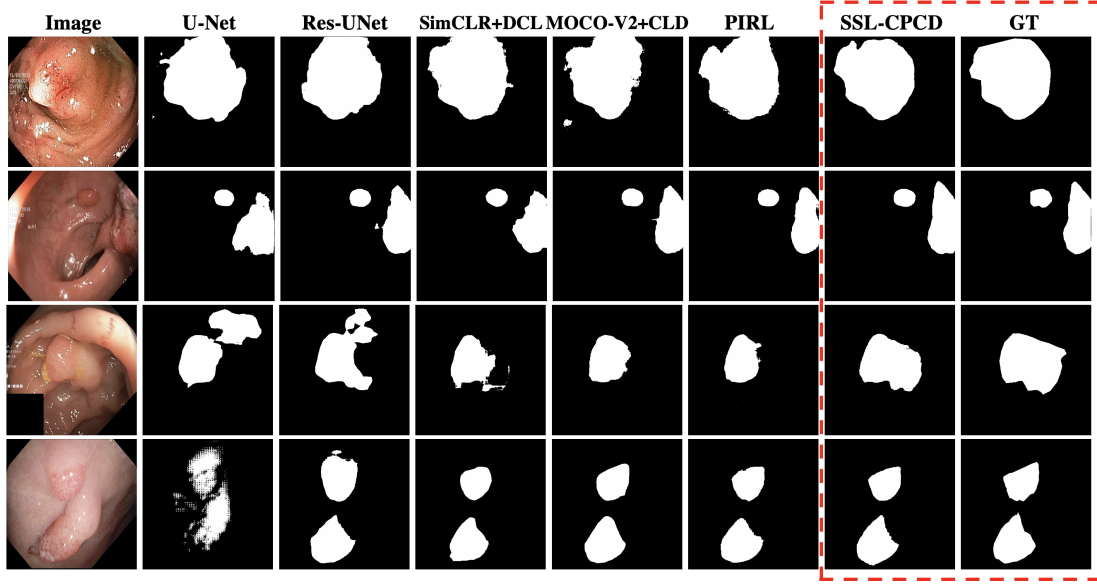


**Figure 5.2:** Qualitative comparison of our proposed self-supervised approach with other state-of-the-art methods for polyp segmentation task.

for main metrics in detection (mAP, Table 5.2) and segmentation (DSC, Table 5.3) is smaller compared to all other methods.

The inevitable camera motion blur as well as lighting artefacts will have an impact on the method. For the detection task in Figure 5.3 (b) one can observe an image in an oblique view that confuses the model and another image with a flat polyp being missed due to specularity surrounding it. A similar under and over-segmentation can be seen for the semantic segmentation task in Figure 5.3 (c).

A similar trend can be observed for detection (Table 5.8) and segmentation (Table 5.9) where nearly 50% labelled data are required by fully supervised methods to reach the performance of SSL-CPCD provided by only 10% of samples. In addition, hidden test centre data (not provided in training) was used to measure the efficacy and robustness of all SSL methods. Similarly with UC classification task, we used a heuristic approach on all SSL approaches to understand the effect of cluster size  $k$  (Table 5.7) and the various other hyperparameters (Table 5.6) in polyp detection and segmentation tasks. The performance of  $k$  is slightly different for different cluster sizes.



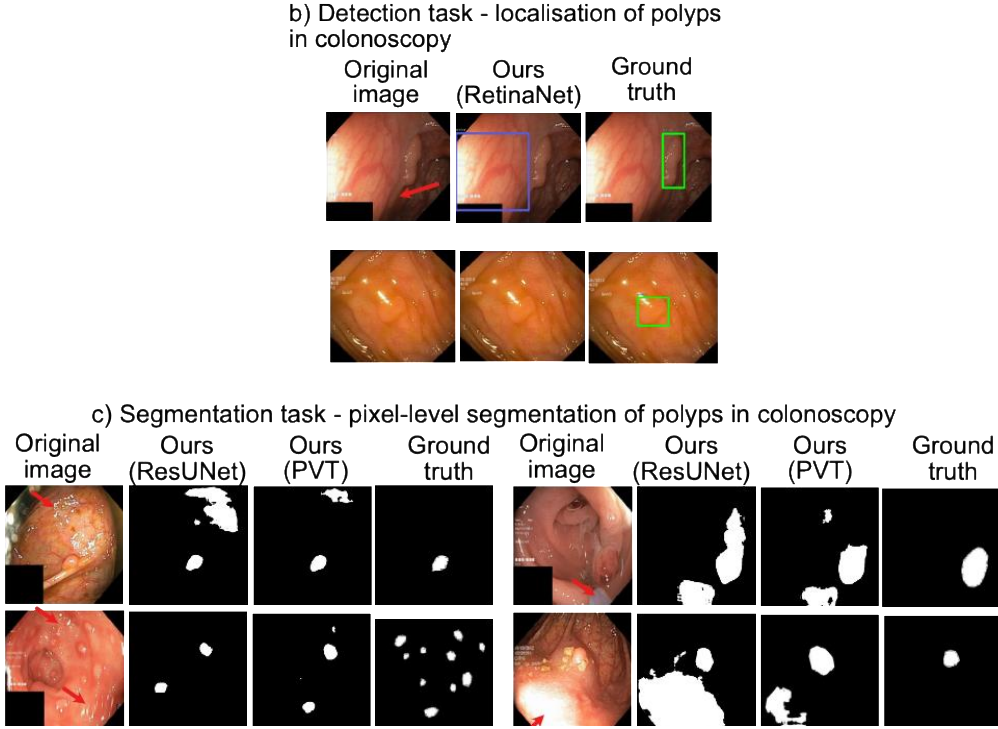


**Figure 5.3:** Limitations of our proposed approach in polyp detection and segmentation tasks.

Our experiments demonstrate that the proposed SSL-CPCD approach outperforms other state-of-the-art self-supervised methods on downstream tasks such as polyp detection and segmentation. The SSL-CPCD framework is able to learn more robust and transferable feature representations from unlabelled data, significantly enhancing the generalisation capacity of the model across different tasks. However, this preliminary investigation also revealed several unmet challenges. Specifically, unavoidable issues such as camera-induced motion blur and illumination artefacts are frequently encountered in real-world video-based polyp segmentation, which are not sufficiently addressed by existing self-supervised methods. These findings motivate the subsequent sections of this chapter, where we propose a more dedicated framework to explicitly account for temporal consistency and visual degradation in video endoscopy, thereby addressing the limitations observed in our earlier experiments.

## 5.3 End-to-end self-supervised and temporal self-attention with feature branching for real-time video polyp segmentation

### 5.3.1 Introduction

Colorectal cancer (CRC) is the third most common cancer worldwide [185]. Early screening and removal of precancerous lesions (colorectal adenomas) can be effective in preventing the development of colorectal cancer and improving survival rates for CRC patients. Evaluation of polyps is an important step in cancer prevention. Several studies [186] provide evidence that survival rates for patients with stage IV and V CRC are reduced by 35% compared to stage I. Colonoscopy has been widely used in clinical practice as the gold standard for CRC screening. During a colonoscopy, an endoscopist confirms the location of rectal polyps and assesses their visual appearance. A comprehensive visual assessment of polyps could eventually reduce the need for unnecessary biopsies, hence reducing the workload for pathologists and impacting the wider patient pathway. However, endoscopy is highly operator-dependent, and many polyps are missed during screening. Computer-assisted approaches are needed to improve diagnostic quality. Here, the accurate segmentation of polyps in real-time is a critical step. Polyps are visually very similar to their surrounding mucosa, and the low contrast affects the effectiveness of most segmentation algorithms. Constant camera motion and artefacts make this task even more challenging. Effective segmentation methods must compensate for motion blur, changes in illumination, occlusion, and specular reflections. While most methods are built upon image-based segmentation techniques [104–106], in the real world, the nature of the colonoscopic procedure is video. Hence, it is crucial to develop real-time video polyp segmentation that leverages the samples from previous frames, tackling false positives. Effectively, this can also improve the robustness of methods when the polyps are occluded or covered with artefacts such as stool. To this end, only a few methods have developed real-time video segmentation. However, this method suffers from generalisability and limits its performance as they are solely

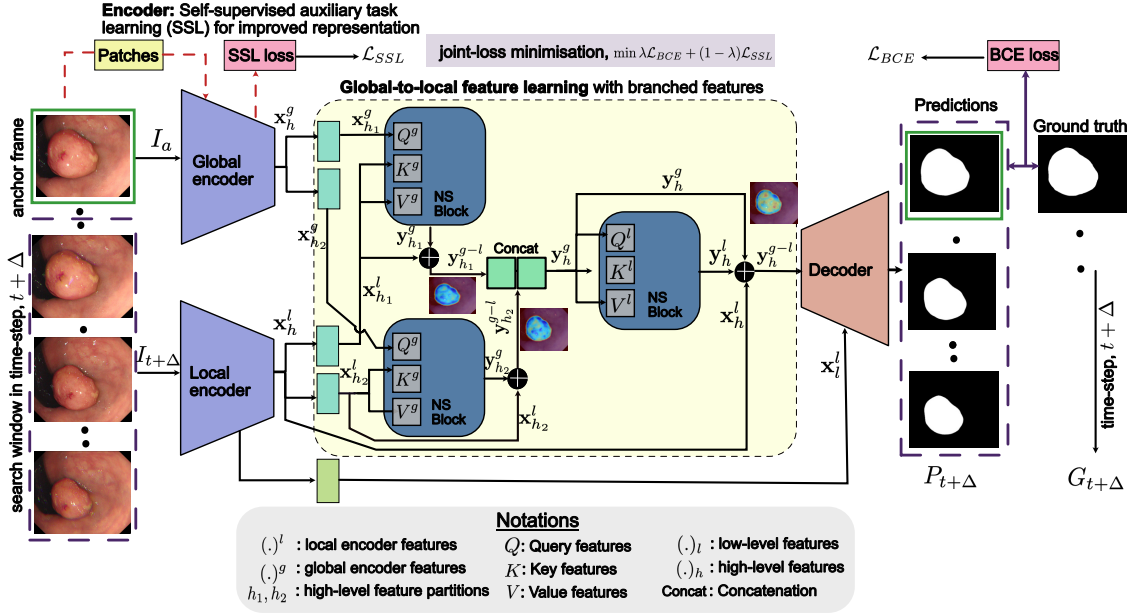


**Figure 5.4:** Proposed network: Our network consists of three parts: a) Global and local encoders trained with a self-supervision loss comprising Jigsaw-puzzle sampling. b) High-level global and local features branch from both encoders and pass through a normalised self-attention block [116], enabling global-to-local feature learning. c) A decoder layer that fuses the self-attention high-level feature maps from NS blocks and the low-level features from the local encoder for final mask prediction.

based on fully supervised learning. Self-supervised learning (SSL) methods [187] have proven to be efficient ways of learning representations by minimising the dissimilarity between the samples. However, the question we have asked is - could we leverage the representation learning as an auxiliary task and incorporate the learnt representations to strengthen the supervised approach? Our work builds upon [117] but incorporates several novel ideas that include - the exploitation of the semantic meaning of frames using an SSL loss within the learning framework, and additionally, we exploit the sub-branching of the feature representations.

In this work, we propose a novel end-to-end self-supervised and normalised self-attention blocks [116] with optimal configurations and feature branching adhering to real-time performance, namely SSTFB network, demonstrating significant improvement over all state-of-the-art (SOTA) methods. Our contributions in this work can be summarised as:

- We propose a self-supervised learning scheme as an auxiliary task in an

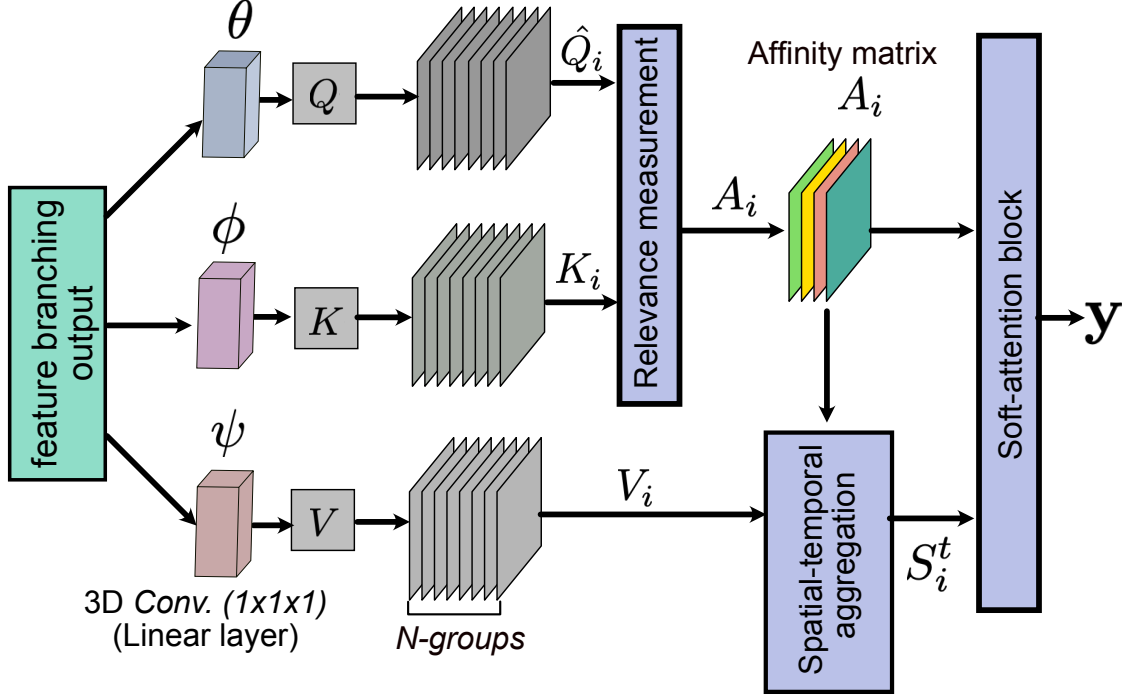


**Figure 5.5:** Normalised self-attention block used in our proposed network.

end-to-end framework to boost feature representations.

- We use feature branching and concatenation techniques for learning separately semantically meaningful features.
- We exploit the spatial-temporal relationship of the features by normalised self-attention blocks with feature aggregation and relevance measurement.
- We show optimal choices of normalised self-attention blocks for global-to-local feature learning without impacting real-time performance.
- We compare on both seen and unseen video datasets with several SOTA methods where our method outperforms them by a larger margin.

### 5.3.2 Methodology

We introduce a joint framework of self-supervised pretext learning and supervised semantic segmentation in video polyp segmentation (Figure 5.4). Our proposed self-supervised auxiliary loss enhances feature representations further in the selected

stream of video samples. Similarly, our supervised self-attention blocks with feature branching mechanism strengthen learning of salient features responsible for polyp segmentation. The entire framework is trained in an end-to-end fashion and does not increase network size making it computationally efficient. Below we describe each blocks separately:

### 5.3.2.1 Encoder self-supervision

We use a pretext-invariant representation learning (PIRL) [120] that uses a Jigsaw puzzle [127] as a pretext task for representation learning. The encoder is fed with the frame sequence ( $\Delta = 5$ ), which is then also used to create patches and shuffled for our auxiliary self-supervised learning (SSL) task. We integrate this task within our proposed segmentation framework. The idea here is to leverage the learnt representations from the SSL towards faster learning and stronger feature representations in our main segmentation task. We do this by jointly minimising the SSL loss and the loss oriented towards our main segmentation task in an end-to-end fashion.

The limitation of SSL is that if training samples are similar, the negative examples used in the NCE loss are more likely to be composed of highly similar instances, especially in our case as we are taking five consecutive frames. Thus, the grouping strategy as shown in our previous work [170] can help mitigate such limitations. A more detailed description of this strategy and its implementation is provided in Section 4.2.2. The idea of patch-level instance group discriminator (PLD) loss [170] is using cross-entropy loss to compute *cosine* similarity between normalised embedding of images  $\bar{r}(\cdot)$  with the centroid of the patch embedding  $r_t(\cdot)$  in a cluster, and normalised embedding of patches  $\bar{r}_t(\cdot)$  with the centroid of the image embedding in a specific cluster. k-means clustering is used to group the feature vectors and compute the cluster centroids for each image embedding  $C^k$  and patch embedding  $PC^k$  where  $k$  is the number of classes. An encoder with free parameters  $\theta$  is used to incarnate representation  $\phi_\theta(\cdot)$ . By assuming the group in

the patch-level cluster is the same as the group in the image-level for that particular class, therefore the PLD loss is defined as:

$$\begin{aligned}\mathcal{L}_{PLD}(\mathbf{I}, \mathbf{I}_t) = & 0.5 \sum_{\forall k} h(\bar{r}(\phi_{\theta}(\mathbf{I})), PC^k) \\ & + 0.5 \sum_{\forall k} h(\bar{r}_t(\phi_{\theta}(\mathbf{I}_t)), C^k).\end{aligned}\quad (5.1)$$

### 5.3.2.2 Normalised self-attention block

A normalised self-attention block as proposed in PNSNet [116] is used. However, unlike [116] we use the feature branches as input to the NS-block, *i.e.* each NS-block receives input  $\mathbf{x} \in \mathcal{R}^{T \times H/2 \times W/2 \times C}$  with temporal frames  $T$ , height  $H$ , width  $W$ , and channel  $C$  (See Fig. 5.5). Three linear embedding functions  $\theta(\cdot)$ ,  $\phi(\cdot)$ , and  $\psi(\cdot)$  are used to generate corresponding attention features using a  $1 \times 1 \times 1$  convolutional layer corresponding to query  $Q$ , key  $K$  and value  $V$ . Each attention feature is then divided into  $N$  groups, and the query features are further normalised for internal covariance shift compensation by utilising layer normalisation. Thus, the set now can be represented as  $\{\hat{Q}_i, K_i, V_i\} \in \mathcal{R}^{T \times H/2 \times W/2 \times C/N}$  with  $i = \{1, \dots, N\}$ . Similar to [116], we used a query-dependent rule where each query pixel  $\mathbf{x}^q$  of  $\hat{Q}_i$  we require to find corresponding key  $K_i$  in a constrained neighbouring window of kernel size  $k$  and dilation rate  $d_i = 2i - 1$  using a sampling function  $\mathcal{F}_s$ . Finally, an affinity matrix  $A_i \in \mathcal{R}^{T \times HW/4 \times T(2k+1)^2}$  is computed as:

$$A_i = \text{Softmax}\left(\frac{\hat{Q}_i \mathcal{F}_s < \mathbf{x}^q, K_i >^T}{\sqrt{C/N}}\right). \quad (5.2)$$

For  $\mathbf{x}^a \in A_i$  we compute inner product with the value  $V_i$  utilising the similar approach as done above with sampling function  $\mathcal{F}_s$  and neighbouring window resulting in a spatial-temporal aggregation matrix,  $S_i^T = A_i \mathcal{F}_s < \mathbf{x}^a, V_i >$ . Finally, a soft-attention is applied to get the final output of the block by concatenating group features along the channel and taking maximum value along these channels for both  $A_i$  and  $S_i^T$ .

### 5.3.2.3 Global-to-local feature learning

Our proposed global-to-local feature learning is unique. Here, we use three NS-blocks exploiting both feature branches from the global encoder for anchor frame  $I_a$  and temporal feature information  $I_{t+\Delta}$  from the local encoder. The global encoder produces features for  $I_a$  extracted from conv4\_6 layer of Res2Net-50[188]. Features extracted for the local encoder for consecutive frames  $I_{t+\Delta}$  is extracted from conv3\_4 and conv4\_6 which are indicated as low-level  $(.)_l$  and high-level  $(.)_h$  features, respectively. Order of feature maps in the branches of high-level features remains the same while entering to two first NS-blocks, i.e., the top feature branch from global encoder  $\mathbf{x}_{h_1}^g$  and the top feature branch from local encoders  $\mathbf{x}_{h_1}^l$  entering the first NS-block while  $\mathbf{x}_{h_2}^g$  and  $\mathbf{x}_{h_2}^l$  are fed to the second NS-block. For arbitrary temporal distance  $\Delta$ , the first two NS-blocks model the long-term relationship, as  $\{\mathbf{x}_l^{h_1}, \mathbf{x}_l^{h_2}\} \in \mathcal{R}^{\Delta \times H/2 \times W/2 \times C}$ . Anchor features are taken as query entries while  $\mathbf{x}_l^{h_1}$  and  $\mathbf{x}_l^{h_2}$  are taken as key and value features. As discussed in Section 5.3.2.2, NS-block creates a spatial-temporal aggregation and pixel-wise similarities between query and key within a neighbour window such that high-level short-term features are exploited. Finally, we use an element-wise addition of features as a residual operation in both NS-blocks, giving  $y_{h_1}^{g-l}$  and  $y_{h_2}^{g-l}$ , respectively. We then concatenate both features from the NS blocks, which is then passed to a third NS block that further enhances the features to propagate into a local neighbourhood, boosting more local representations and two residual connections to avoid the vanishing gradient problem. It can be observed that for  $y_h^{g-l}$ , both local and global features are pronounced after performing these operations and hold spatial-temporal information.

### 5.3.2.4 Decoder

Learned features from the global-to-local feature block that consists of the high-level spatial-temporal features  $y_h^{g-l}$  together with the low-level features from the local encoder for time frame  $t$  represented as  $\mathbf{x}_l^l$  is fed to a two-stage UNet-like decoder  $D$  estimating predictions  $P_t$ . These predictions are compared with ground truth using widely used binary cross-entropy  $\mathcal{L}_{\text{BCE}}$ .

### 5.3.2.5 Joint-loss optimisation (end-to-end framework)

Our novel approach includes joint-loss optimisation for both auxiliary tasks, i.e., self-supervised pretext learning (SSL) and the main fully supervised segmentation task, i.e., segmentation in an end-to-end fashion. The SSL loss  $\mathcal{L}_{SSL}$  is given by combining the noise contrastive loss ( $\mathcal{L}_{NCE}$ ) and the patch-level group discrimination loss ( $\mathcal{L}_{PLD}$ ) described in Section 5.3.2.1. Our experiments suggests equal weighting of these two losses thus the final SSL loss can be represented as:

$$\mathcal{L}_{SSL} = 0.5 \times \mathcal{L}_{NCE} + 0.5 \times \mathcal{L}_{PLD} \quad (5.3)$$

For the segmentation task, we utilise predicted segmentation mask  $P_t$  and ground-truth mask  $G_t$  at the same time frame  $t$  to minimise a binary cross-entropy (BCE) loss formulated as:

$$\mathcal{L}_{BCE} = - \sum (G_t \log(P_t) + (1 - G_t) \log(1 - P_t)) \quad (5.4)$$

Thus, the joint-loss  $\mathcal{L}_{\text{joint}}$  is minimised by a combination of SSL loss and segmentation loss. Thus, the final loss with  $\lambda$  as weighting factor,  $\mathcal{L}_{\text{joint}}$  is given by:

$$\mathcal{L}_{\text{joint}} = \lambda \mathcal{L}_{BCE} + (1 - \lambda) \mathcal{L}_{SSL} \quad (5.5)$$

In our experiments we empirically set  $\lambda = 0.5$  for the joint loss.

## 5.3.3 Experimental setup

### 5.3.3.1 Datasets and Comparison models

Experiments were conducted on two widely used video polyp segmentation datasets: CVC-612 [173] and SUN-SEG [183]. SUN-SEG is a large colonoscopy video database from Showa University and Nagoya University. The dataset comprises 49,136 polyp frames extracted from 100 polyp videos and the corresponding masks. We split 40% of SUN-SEG data for training and the rest of the data for testing. The test data are divided into SUN-SEG-Easy (seen & unseen) and SUN-SEG-Hard (seen & unseen) according to the polyp pathology’s difficulty level. Seen indicates that the testing samples are from the same case in the training set, whereas the



unseen denotes that samples do not exist in the training set. We compare our approach with ten most recent polyp segmentation methods: PraNet [105], Polyp-PVT [106], SANet [189], PNSNet [116], TGANet [190], PNS+ [117], CFANet [113], ICBNet [119], PVT-CASCADE [107] and CTNet [108].

### 5.3.3.2 Evaluation metrics

Two conventional metrics, mean Dice Similarity Coefficient ( $mDice = \frac{2TP}{2TP+FP+FN}$ ) and mean Intersection over Union ( $mIoU = \frac{TP}{TP+FP+FN}$ ) to measure the similarity and overlap between predictions and ground truth are used. Here, TP is true positives, FP is false positives, and FN refers to the false negatives.

In addition, we have also included the S-measure, which is the weighted ( $\alpha$ ) metric between the object-aware structural similarity measure ( $S_o$ ) and region-aware structural similarity measure ( $S_r$ ) [191] ( $S_\alpha = \alpha \times S_o(P, G) + (1 - \alpha) \times S_r(P, G)$ ). Here,  $P$  is prediction,  $G$  is ground true, and  $\alpha$  is set to 0.5.

A weighted F-measure which corrects for “Equal-importance flaws” in Dice is also used [192, 193] ( $F_\beta^w = (1 + \beta^2) \times precision^w \times recall^w / (\beta^2 \times (precision^w + recall^w))$ ). Here,  $precision = \frac{TP}{TP+FP}$ ,  $recall = \frac{TP}{TP+FN}$ , and  $\beta^2 = 0.3$ .

Another metric referred to as an enhanced-alignment measure which is suitable for assessing heterogeneous location and shape of polyps in colonoscopy [117] is also used ( $E_\phi = (1/(W \times H)) \sum_{x=0}^W \sum_{y=0}^H \phi(P(x, y), G(x, y))$ ,  $W$ : width of  $G$ ,  $H$ : height of  $G$ ,  $\phi$ : enhanced alignment matrix) [194].

### 5.3.3.3 Implementation details

We train our SSTFB network using the PyTorch framework on a server deployed with NVIDIA Quadro RTX 6000 graphics cards. All input images were resized to  $256 \times 448$  pixels, and batch size was set to 24. The first video clip frame was chosen as an anchor and randomly sampled five consecutive frames as neighbouring frames ( $\Delta = 5$ ). Throughout the model training process, we set the initial learning rate and weight decay of the Adam optimiser to  $3e^{-4}$  and  $1e^{-4}$ , respectively. The number of stacked normalised self-attention blocks is set as  $R = 3$ , the kernel size  $k = 3$ , and the number of attention groups to  $N = 4$  with dilation rate of  $\{3, 4, 3, 4\}$

for the first two NS-blocks, and dilation rate of  $\{1, 2, 1, 2\}$  for the third NS-block. For the SSL task, we use the patches from the entire sequence of images fed to the Res2Net-50 model (pre-trained with ImageNet weights).

### 5.3.4 Results

#### 5.3.4.1 Method comparisons

In this section, we assess the performance of our method on two seen test (held-out) datasets (SUN-SEG-Easy(seen) & SUN-SEG-Hard(seen)) and comparison with other SOTA methods.

#### Quantitative results

Quantitative results of SOTA methods and our proposed SSTFB are summarised in Table 5.11. It can be observed from Table 5.11 that the performance of the method developed for still polyp image segmentation (PraNet, Polyp-PVT, SANet, TGANet, CFANet, PVT-CASCADE and CTNet) is lower than the video polyp image segmentation method (PNSNet, PNS+ ,ICBNet and SSTFB(ours)) on both SUN-SEG-Easy and SUN-SEG-Hard datasets. Specifically, our proposed SSTFB reaches 0.911 on mDice, 0.856 on mIoU,  $S_\alpha$  on 0.935,  $E_\phi$  on 0.968 and  $F_\beta^w$  on 0.871 achieving performance improvements of 9.9%, 11.9%, 7.8%, 7.4% and 12.8% respectively compared with the polyp segmentation benchmark work (PraNet). SSTFB also achieves 4.4% and 4.9% increase of mDice and mIoU on SUN-SEG-Easy(seen) when contrasted with the recent SOTA method (CTNet). Simultaneously, our method still has a mDice improvement of about 3% for the most recent SOTA polyp video segmentation model PNS+. On the SUN-SEG-Hard(seen) dataset, SSTFB presents mDice of 0.881 and mIoU of 0.813, which shows an over 8% performance improvement compared with PraNet and 2.6% improvement compared with PNS+.

Table 5.13 shows SSTFB can run at extremely high inference speeds (126fps). Our proposed SSTFB significantly outperforms all other SOTA models on all metrics with real-time performance. This shows that our model has a stronger

**Table 5.11:** Quantitative results on seen test datasets

| Method            | Publication            | SUN-SEG-Easy (seen) |              |              |              | SUN-SEG-Hard (seen) |              |              |              |              |              |
|-------------------|------------------------|---------------------|--------------|--------------|--------------|---------------------|--------------|--------------|--------------|--------------|--------------|
|                   |                        | mDice               | mIoU         | $S_\alpha$   | $E_\phi$     | $F^w_\beta$         | mDice        | mIoU         | $S_\alpha$   | $E_\phi$     | $F^w_\beta$  |
| PraNet [105]      | MICCAI <sub>20</sub>   | 0.812               | 0.737        | 0.857        | 0.894        | 0.743               | 0.795        | 0.721        | 0.834        | 0.860        | 0.736        |
| Polyp-PVT [106]   | MICCAI <sub>21</sub>   | 0.819               | 0.740        | 0.862        | 0.899        | 0.748               | 0.801        | 0.727        | 0.841        | 0.862        | 0.747        |
| SANet [189]       | MICCAI <sub>21</sub>   | 0.827               | 0.743        | 0.865        | 0.909        | 0.751               | 0.807        | 0.732        | 0.845        | 0.866        | 0.751        |
| TGANet [190]      | MICCAI <sub>22</sub>   | 0.833               | 0.752        | 0.876        | 0.915        | 0.76                | 0.814        | 0.739        | 0.851        | 0.871        | 0.755        |
| PNSNet [116]      | MICCAI <sub>21</sub>   | 0.835               | 0.760        | 0.882        | 0.922        | 0.771               | 0.821        | 0.753        | 0.860        | 0.882        | 0.769        |
| PNS+ [117]        | MIR <sub>22</sub>      | 0.887               | 0.826        | 0.915        | 0.954        | 0.844               | 0.855        | 0.787        | 0.885        | 0.927        | 0.796        |
| CFANet [113]      | PR <sub>23</sub>       | 0.841               | 0.767        | 0.88         | 0.88         | 0.785               | 0.829        | 0.757        | 0.866        | 0.866        | 0.762        |
| ICBNet [119]      | BIBM <sub>23</sub>     | 0.846               | 0.775        | 0.887        | 0.921        | 0.781               | 0.827        | 0.755        | 0.862        | 0.887        | 0.767        |
| PVT-CASCADE [107] | WACV <sub>23</sub>     | 0.855               | 0.789        | 0.789        | 0.93         | 0.792               | 0.832        | 0.762        | 0.869        | 0.891        | 0.783        |
| CTNet [108]       | IEEE TOC <sub>24</sub> | 0.867               | 0.807        | 0.899        | 0.939        | 0.816               | 0.84         | 0.769        | 0.873        | 0.895        | 0.787        |
| SSTFB             | Ours <sub>24</sub>     | <b>0.911</b>        | <b>0.856</b> | <b>0.935</b> | <b>0.968</b> | <b>0.871</b>        | <b>0.881</b> | <b>0.813</b> | <b>0.909</b> | <b>0.942</b> | <b>0.829</b> |

**Table 5.12:** Quantitative results on unseen test datasets

| Method            | Publication            | SUN-SEG-Easy (unseen) |              |              |              | SUN-SEG-Hard (unseen) |              |              |              |              |              |
|-------------------|------------------------|-----------------------|--------------|--------------|--------------|-----------------------|--------------|--------------|--------------|--------------|--------------|
|                   |                        | mDice                 | mIoU         | $S_\alpha$   | $E'_\phi$    | $F^w_\beta$           | mDice        | mIoU         | $S_\alpha$   | $E'_\phi$    | $F^w_\beta$  |
| PraNet [105]      | MICCAI <sub>20</sub>   | 0.669                 | 0.581        | 0.727        | 0.754        | 0.629                 | 0.651        | 0.572        | 0.719        | 0.741        | 0.601        |
| Polyp-PVT [106]   | MICCAI <sub>21</sub>   | 0.675                 | 0.589        | 0.737        | 0.760        | 0.641                 | 0.663        | 0.592        | 0.706        | 0.737        | 0.593        |
| SANet [189]       | MICCAI <sub>21</sub>   | 0.687                 | 0.594        | 0.723        | 0.762        | 0.645                 | 0.654        | 0.565        | 0.713        | 0.745        | 0.617        |
| TGANet [190]      | MICCAI <sub>22</sub>   | 0.672                 | 0.577        | 0.753        | 0.759        | 0.631                 | 0.665        | 0.577        | 0.749        | 0.768        | 0.638        |
| PNSNet [116]      | MICCAI <sub>21</sub>   | 0.703                 | 0.619        | 0.764        | 0.778        | 0.653                 | 0.676        | 0.581        | 0.765        | 0.776        | 0.651        |
| PNS+ [117]        | MIR <sub>22</sub>      | 0.749                 | 0.652        | 0.793        | 0.829        | 0.711                 | 0.736        | 0.639        | 0.795        | 0.815        | 0.704        |
| CFANet [113]      | PR <sub>23</sub>       | 0.701                 | 0.607        | 0.76         | 0.771        | 0.649                 | 0.672        | 0.579        | 0.758        | 0.769        | 0.645        |
| ICBNet [119]      | BIBM <sub>23</sub>     | 0.717                 | 0.615        | 0.769        | 0.782        | 0.661                 | 0.683        | 0.595        | 0.771        | 0.779        | 0.657        |
| PVT-CASCADE [107] | WACV <sub>23</sub>     | 0.721                 | 0.619        | 0.771        | 0.788        | 0.672                 | 0.689        | 0.607        | 0.774        | 0.78         | 0.662        |
| CTNet [108]       | IEEE TOC <sub>24</sub> | 0.726                 | 0.635        | 0.778        | 0.79         | 0.676                 | 0.717        | 0.618        | 0.782        | 0.796        | 0.68         |
| SSTFB             | Ours <sub>24</sub>     | <b>0.772</b>          | <b>0.687</b> | <b>0.846</b> | <b>0.877</b> | <b>0.748</b>          | <b>0.767</b> | <b>0.673</b> | <b>0.822</b> | <b>0.843</b> | <b>0.728</b> |

learning ability and can effectively learn semantically meaningful features for polyp segmentation.

**Table 5.13:** Trainable parameters and frames-per-second (FPS)

| Method       | Trainable parameters | FPS    |
|--------------|----------------------|--------|
| PraNet       | 39.2M                | 97fps  |
| Polyp-PVT    | 125.6M               | 66fps  |
| SANet        | 45.1M                | 72fps  |
| TGANet       | 42.3M                | 85fps  |
| PNSNet       | 29.7M                | 135fps |
| PNS+         | 26.8M                | 140fps |
| CFANet       | 25.2M                | 86fps  |
| ICBNet       | 62.1M                | 71fps  |
| PVT-CASCADE  | 134.5M               | 62fps  |
| CTNet        | 168.9M               | 57fps  |
| SSTFB (ours) | 33.4M                | 126fps |

#### 5.3.4.2 Generalization

For the generalisability study, we have trained our method and other SOTA methods on the SUN-SEG dataset and experimented on two new unseen and one out-of-centre polyp video datasets, namely, SUN-SEG-Easy(unseen), SUN-SEG-Hard(unseen) and CVC-612.

##### Generalizability results on unseen datasets

As can be seen from Table 5.12, our method is consistently superior to all SOTA methods on SUN-SEG-Easy(unseen) dataset(e.g., nearly 5%-10% higher value on mDice and other metrics compared to PraNet, Polyp-PVT, SANet, TGANet, CFANet, ICBNet, PVT-CASCADE and CTNet). Compared to SOTA video polyp segmentation methods (PNS+), our approach still shows over 3% improvement in mDice, mIoU and  $S_\alpha$ . Especially in SUN-SEG-Hard(unseen) dataset, SSTFB still maintains competitive performance of 0.767, 0.673, 0.822, 0.843 and 0.728 on mDice, mIoU,  $S_\alpha$ ,  $E_\phi$  and  $F_\beta^w$ , respectively. The performance of most methods drops sharply while our method retains its performance with performance improvement even over 10% compared to most SOTA methods.

**Table 5.14:** Quantitative results on out-of-centre test dataset

| Method            | Publication            | CVC-612      |              |              |              |              |
|-------------------|------------------------|--------------|--------------|--------------|--------------|--------------|
|                   |                        | mDice        | mIoU         | $S_\alpha$   | $E_\phi$     | $F_\beta^w$  |
| PraNet [105]      | MICCAI <sub>20</sub>   | 0.817        | 0.733        | 0.856        | 0.885        | 0.752        |
| Polyp-PVT [106]   | MICCAI <sub>21</sub>   | 0.826        | 0.744        | 0.861        | 0.891        | 0.769        |
| SANet [189]       | MICCAI <sub>21</sub>   | 0.819        | 0.738        | 0.849        | 0.889        | 0.763        |
| TGANet [190]      | MICCAI <sub>22</sub>   | 0.832        | 0.751        | 0.879        | 0.896        | 0.775        |
| PNSNet [116]      | MICCAI <sub>21</sub>   | 0.822        | 0.739        | 0.857        | 0.883        | 0.752        |
| PNS+ [117]        | MIR <sub>22</sub>      | 0.841        | 0.763        | 0.882        | 0.907        | 0.793        |
| CFANet [113]      | PR <sub>23</sub>       | 0.82         | 0.734        | 0.855        | 0.886        | 0.771        |
| ICBNet [119]      | BIBM <sub>23</sub>     | 0.835        | 0.755        | 0.871        | 0.889        | 0.767        |
| PVT-CASCADE [107] | WACV <sub>23</sub>     | 0.837        | 0.76         | 0.878        | 0.899        | 0.772        |
| CTNet [108]       | IEEE TOC <sub>24</sub> | 0.844        | 0.771        | 0.889        | 0.905        | 0.796        |
| SSTFB             | Ours <sub>24</sub>     | <b>0.862</b> | <b>0.786</b> | <b>0.893</b> | <b>0.923</b> | <b>0.814</b> |

**Table 5.15:** Ablation experiments of SSTFB on different datasets

| Type                     | Easy(seen)   |              | Hard(seen)   |              | Easy(unseen) |              | Hard(unseen) |              | CVC-612      |              |
|--------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                          | mDice        | $S_\alpha$   | mDice        | $S_\alpha$   | mDice        | $S_\alpha$   | mDice        | $S_\alpha$   | mDice        | $S_\alpha$   |
| 3NS-block w/o SSL        | 0.931        | 0.937        | 0.895        | 0.913        | 0.822        | 0.874        | 0.791        | 0.839        | 0.877        | 0.919        |
| 3NS-block + SSL w/o e2e  | 0.936        | 0.947        | 0.897        | 0.921        | 0.816        | 0.871        | 0.806        | 0.854        | 0.885        | 0.927        |
| 3NS-block + SSL in e2e   | 0.943        | 0.950        | 0.902        | 0.929        | 0.829        | 0.878        | 0.815        | 0.866        | 0.903        | 0.931        |
| 3NS-block + SSL in e2e+b | <b>0.959</b> | <b>0.962</b> | <b>0.925</b> | <b>0.938</b> | <b>0.841</b> | <b>0.899</b> | <b>0.827</b> | <b>0.871</b> | <b>0.911</b> | <b>0.942</b> |

### Generalizability results on out-of-centre datasets

We also evaluated our method against out-of-centre data in Table 5.14 where we can observe that our proposed method achieves a mDice of 0.862, mIoU of 0.786,  $S_\alpha$  of 0.893,  $E_\phi$  of 0.923 and  $F_\beta^w$  of 0.814, respectively. Compared to PraNet and PNS+, our approach exhibits over 4% and 2% enhancement in mDice. Overall, the proposed SSTFB achieves superior performance compared to all other SOTA methods.

**Table 5.16:** Ablation with different NS-blocks in SSTFB

| Type               | Easy(seen)   |              | Hard(seen)   |              | Easy(unseen) |              | Hard(unseen) |              | CVC-612      |              |
|--------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                    | mDice        | $S_\alpha$   | mDice        | $S_\alpha$   | mDice        | $S_\alpha$   | mDice        | $S_\alpha$   | mDice        | $S_\alpha$   |
| 2 NS-blocks        | 0.944        | 0.950        | 0.909        | 0.921        | 0.836        | 0.885        | 0.816        | 0.857        | 0.899        | 0.932        |
| 3 NS-blocks (ours) | <b>0.959</b> | <b>0.962</b> | <b>0.925</b> | <b>0.938</b> | <b>0.841</b> | <b>0.899</b> | <b>0.827</b> | <b>0.871</b> | <b>0.911</b> | <b>0.942</b> |
| 4 NS-blocks        | 0.951        | 0.957        | 0.916        | 0.929        | 0.832        | 0.881        | 0.820        | 0.864        | 0.906        | 0.937        |

**Table 5.17:** Paired  $t$ -test on different datasets. p-values (cutoff at  $\geq 0.05$ ) demonstrates significant difference between SSTFB and other methods.

| comparison        | Easy(seen) | Hard(seen) | Easy(unseen) | Hard(unseen) | CVC-612 |
|-------------------|------------|------------|--------------|--------------|---------|
|                   | p-value    | p-value    | p-value      | p-value      | p-value |
| SSTFB & PraNet    | 9.71E-14   | 8.03E-09   | 0.036        | 0.0195       | 0.0327  |
| SSTFB & Polyp-PVT | 7.59E-19   | 2.17E-10   | 0.0148       | 0.0096       | 0.0219  |
| SSTFB & PNS+      | 3.91E-56   | 6.46E-11   | 0.0077       | 0.00042      | 0.0092  |

#### 5.3.4.3 Ablation Study

We have conducted an extensive ablation study on five different datasets (SUN-SEG-Easy(seen), SUN-SEG-Hard(seen), SUN-SEG-Easy(unseen), SUN-SEG-Hard(unseen), and CVC-612) to evaluate the effectiveness of our proposed method. First, we ablate the impact of SSL block and feature branching. Moreover, we will conduct an extended ablation experiment to further evaluate our approach’s performance under different NS-Blocks settings.

**Effectiveness of SSL block:** The external SSL block that uses SSL pre-training and the built-in SSL block in our end-to-end model are tested separately. From Table 5.15, we can observe that the end-to-end (e2e) model with built-in SSL block achieves 0.7%, 0.5%, 1.3%, 0.9% and 1.8% of mDice improvement on five datasets. Meanwhile, the e2e model is also better than the external model (about 1% improvement) in  $S_\alpha$ . This performance improvement shows that the joint training of the entire model through the end-to-end SSL block training can help the model learn to capture semantically meaningful polyp features.

**Effectiveness of feature branching:** In Table 5.15, the last row in the table represents the method’s performance after using feature branching (SSL+e2e+b). Compared with other methods that are without feature branching, reported results have a significant performance improvement. Specifically, removal of feature branching shows a decrease in mice by 1.6%, 2.3%, 1.2%, 1.2% and 0.8% and reductions in  $S_\alpha$  by 1.2%, 0.9%, 2.1%, 0.5% and 1.1% for SEG-Easy(seen), SUN-SEG-Hard(seen), SEG-Easy(unseen), SUN-SEG-Hard(unseen) and CVC-612. The results show that synthesising aggregated features by branching global and local features and input into soft attention blocks separately improves performance.

**Choice of NS-block:** The choice of the number of NS blocks with feature branching and fusion was also tested with our experiments demonstrating the current configuration to be the most optimal (See Table 5.16). As shown in Table 5.16, our method with three NS-blocks settings demonstrates superior performance, which has approximately a 1.5% increase in both mDice and  $S_\alpha$  on all datasets. The model’s performance with four NS-blocks is slightly higher than with the two NS-blocks

setting.

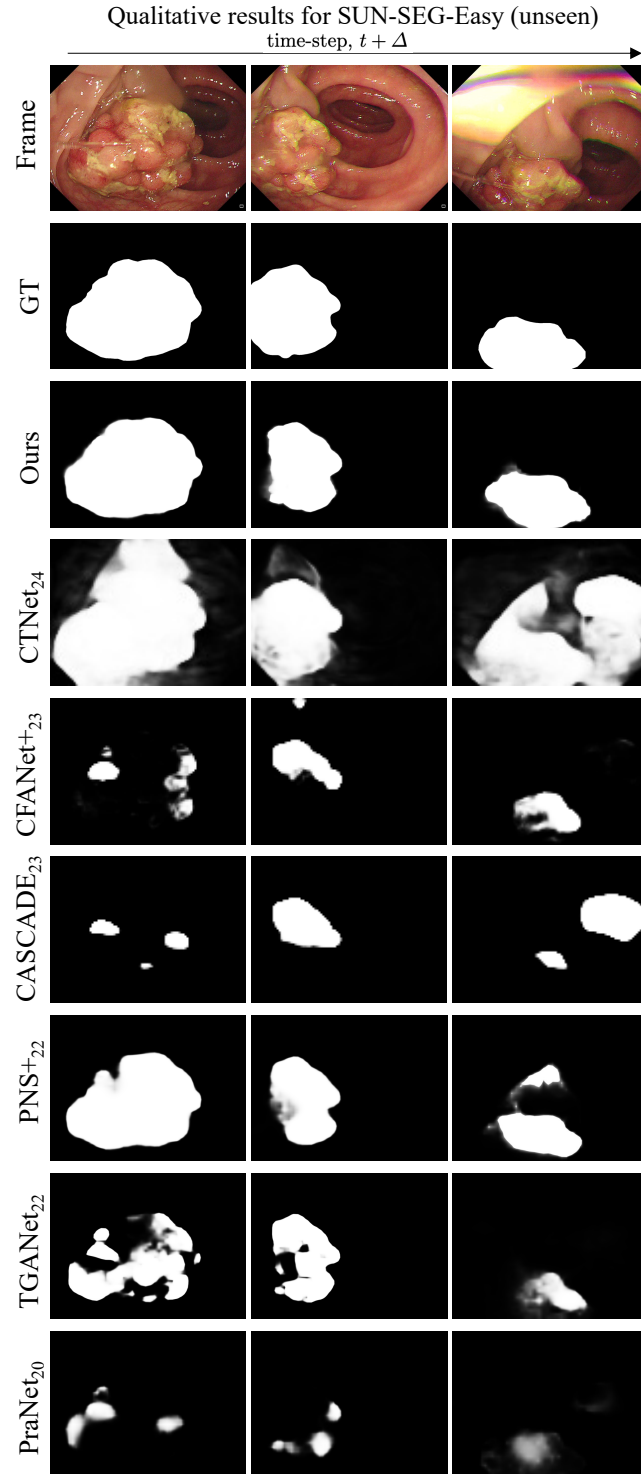


**Figure 5.6:** Qualitative results: Easy samples from SUN-SEG-Easy (unseen).

#### 5.3.4.4 Qualitative results

Fig. 5.6 and Fig. 5.7 demonstrate the effectiveness of our approach and competing networks for the video polyp segmentation on SUN-SEG and CVC-612 datasets. It can be observed that methods designed based on the video polyp segmentation task (SSTFB (Ours) and PNS+) show more obvious performances compared with methods applied to still polyp segmentation (PraNet and TGANet). Specifically, our proposed method demonstrates superior performance with sharper boundaries when dealing with multi-scale polyp segmentation and outperforms competing networks on various challenging video polyp segmentation tasks. For example, in the first to third columns of Fig. 5.6, competing methods vaguely locate polyps to a large extent, while SSTFB can accurately locate polyps. In the SUN-SEG-Hard part of Fig. 5.7, most competing methods cannot accurately distinguish polyps from similar environments, which results in blurred edges of predicted polyps. Our method can cope with background interference and produce accurate edge information. Bubbles on the lens will also affect the judgment of the method to a certain extent. In the last row of the CVC-612 part of Fig. 5.7, the competing method was affected by the bubbles and incorrectly judged the location of the polyp. Nonetheless, our method can still ignore the effect of bubbles and provide accurate segmentation results. The proposed SSTFB demonstrates a competitive advantage over other state-of-the-art networks in video polyp segmentation tasks.

#### 5.3.5 Discussion

Due to the similar environment of the internal structure of the intestine, frequent movement of the endoscope, and lighting issues, there are inevitable problems with similar disease areas to the intestinal wall in endoscopy videos, artefacts, saturation, and motion camera blur. Some images with lower scores are presented in Fig. 5.8. For example, in the first and third rows, the method is affected by flares and pixel saturation and incorrectly determines the polyp area. In addition, air bubbles (second row) and surgical instruments (fourth row) also cause false positive segmentation results. Additionally, the third row is affected



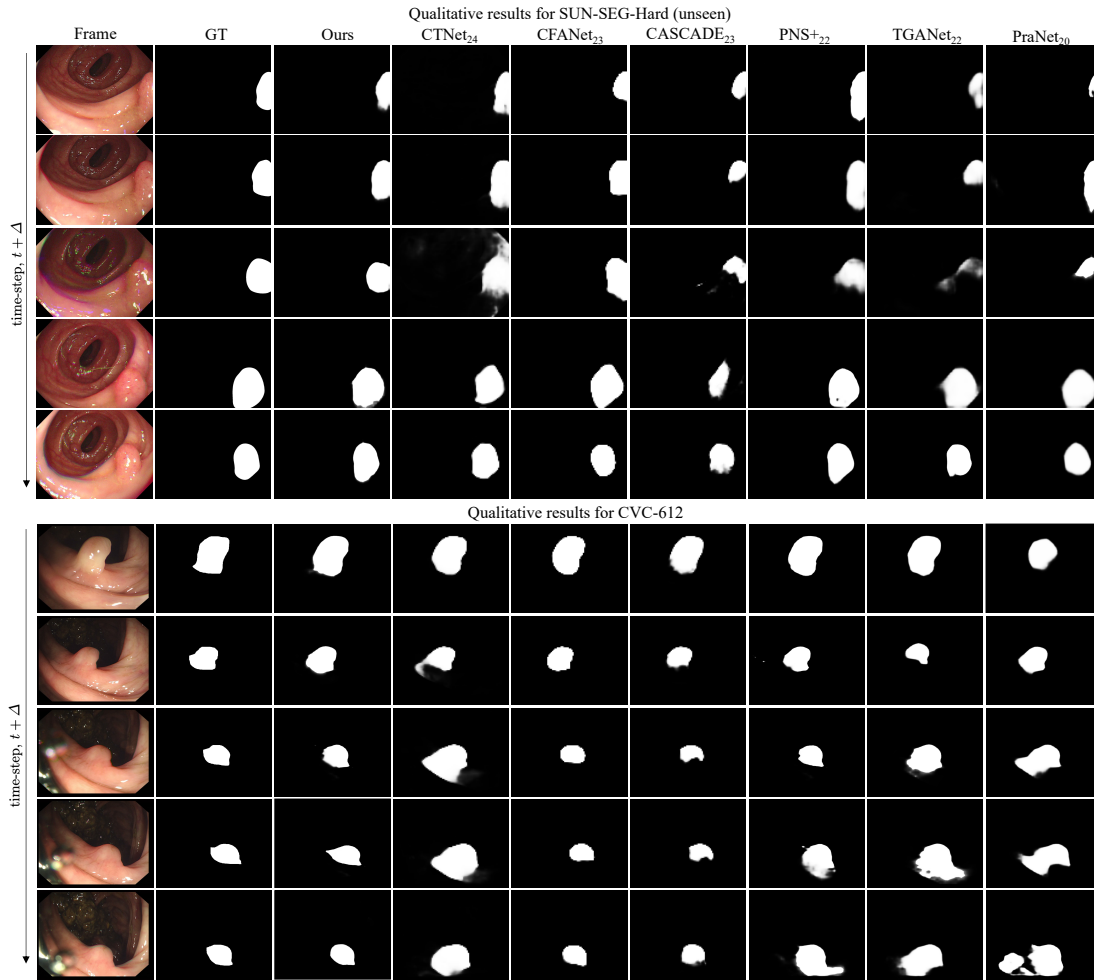


**Figure 5.7:** Qualitative results: Hard samples from SUN-SEG-Hard (unseen, top rows), and samples from unseen data centre CVC-612 (bottom rows).

by various colour artefacts (chromatic aberration) due to sudden motion changes or reflection. In these cases, a depressed region with such chromatic aberration is determined falsely as polyps. Mitigating data bias due to such variability as the causal effect in video data is important and can be tackled by video-based polyp segmentation only to some extent (Fig. 5.6 and Fig. 5.7), but it still outputs some false positives (Fig. 5.8). Compared to other methodologies developed based on static polyp images, approaches incorporating normalised self-attention blocks and a global-to-local learning strategy can more accurately segment polyp positions and contour information by leveraging both long-term and short-term spatial-temporal cues. Furthermore, integrating embedded SSL and sub-branching mechanisms can further learn meaningful representations and enhance the feature-separation

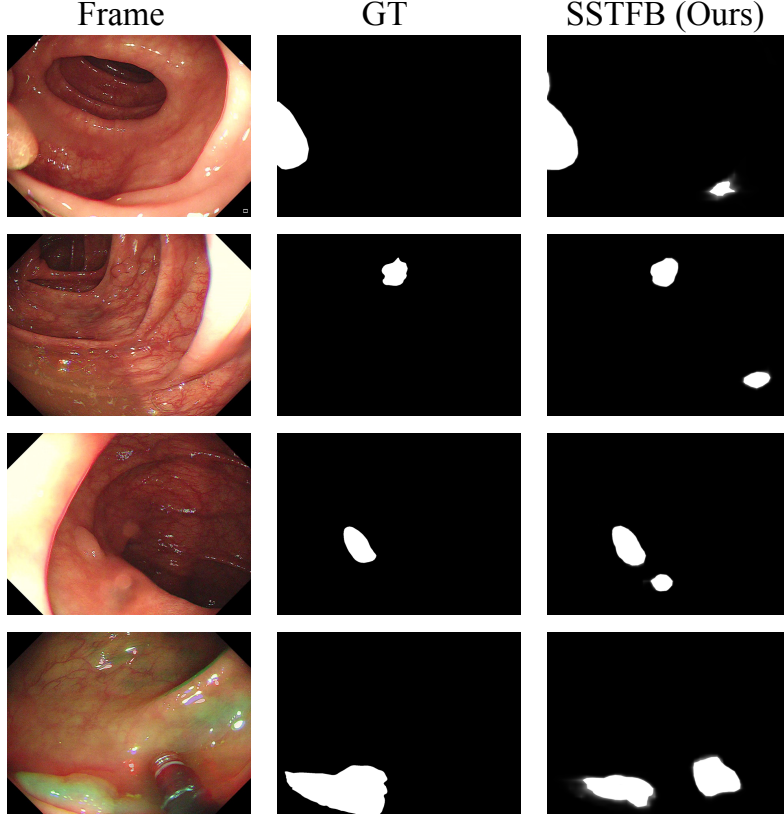


**Figure 5.8:** Limitations of our proposed approach, including bubbles, instruments, and imaging artefacts such as specularity and pixel saturation.

capability of the normalised self-attention blocks. SSTFB has demonstrated superior segmentation results within both SUN-SEG-Easy(Fig. 5.6), SUN-SEG-Hard and CVC-612(Fig. 5.7) datasets.

Furthermore, p-values from paired  $t$ -tests represent significant differences between SSTFB and other approaches. It can be observed from Table 5.17 that our approach provided p-value  $\leq 0.05$  (significantly different) compared with PraNet, Polyp-PVT and PNS+ in all five different datasets. Our network ablation on utilising an optimal number of NS blocks (Fig. 5.16) also demonstrates that the chosen architecture is best suited for video polyp segmentation in contrast to other works previously published.

Similarly, end-to-end learning of the video polyp segmentation, jointly fine-tuned with learned feature representation, SSL, and feature branching step, thereby segregating high-level and low-level features and their components, made the network robust to seen, unseen and out-of-distribution datasets (Table 5.15). It can be

observed that performing self-supervised learning (pretext task for representation learning) does improve the result, but using the proposed combination in an end-to-end fashion favours learning. This also lowers the bottleneck of computational time requirement during training and provides more useful feature learning. As evident from our experiments, the proposed SSTFB has real-time performance (126FPS, Table 5.13) with accuracies higher than all SOTA methods for polyp segmentation in the literature on SUN-SEG easy and hard samples. The improvement on seen (Table 5.11) and unseen (Table 5.12) samples ranges from nearly 10% (PraNet) up to nearly 3% (PNS+) on mDice metric.

### 5.3.6 Conclusion

We proposed a novel end-to-end framework exploiting spatial-temporal information, representation learning using self-supervision, and feature branching with normalised self-attention blocks, namely SSTFB. We provided detailed ablation of various network configurations that suggest that the proposed network outperforms existing SOTA methods without hurting performance time. Compared to many methods in SOTA that are trained and validated on single-frame endoscopy datasets, we demonstrated that our approach outperformed these methods by a larger margin and hence can be used for precise and real-time polyp segmentation in colonoscopy videos that are usually of natural occurrences in clinical settings. The proposed SSL block integration and the normalised self-attention block based on feature branching presented in this chapter can be flexibly used in existing CNN-based segmentation models. Our extensive experiments showed the superiority of our method over all existing SOTA methods in all evaluation metrics, generalisability assessment, and visual quantification.

# 6

## Conclusion and Future work

### Contents

---

|            |                                                             |            |
|------------|-------------------------------------------------------------|------------|
| <b>6.1</b> | <b>Summary of Contributions . . . . .</b>                   | <b>126</b> |
| <b>6.2</b> | <b>Limitations and directions for future work . . . . .</b> | <b>128</b> |
| 6.2.1      | Image quality enhancement based on artefact restoration     | 128        |
| 6.2.2      | More comprehensive grading strategy for ulcerative colitis  | 129        |
| 6.2.3      | Quantification of disease burden . . . . .                  | 130        |
| <b>6.3</b> | <b>Conclusion . . . . .</b>                                 | <b>131</b> |

---

This thesis investigates and evaluates a variety of AI-based image analysis methods, aiming to turn AI into an auxiliary tool for endoscopists to achieve automatic and accurate endoscopy. This chapter provides a summary of the key contributions of this thesis, examines the limitations of the present study, and outlines possible avenues for future research.

### 6.1 Summary of Contributions

This thesis addresses several specific research questions in endoscopic surgery. The following major contributions were made during the completion process:

- **Chapter 1** of this thesis introduces the process of endoscopy and explains the motivation for developing computer vision algorithms to assist endoscopists in performing endoscopy.

- **Chapter 2** of this thesis presents the clinical background of endoscopy, specifically addressing inflammatory bowel disease and polyps. In the technical background, the existing image classification, detection and segmentation methods are comprehensively reviewed, emphasizing the advantages of deep learning in image analysis. Consequently, this underscores the importance of advancing deep learning-based approaches for endoscopic applications.
- **Chapter 3** of this thesis presents an image quality assessment system based on artefact detection. The proposed artefact detection method is a model scaling network based on a novel bounding box pruning approach. The proposed pruning-based mechanism is a post-processing method that prunes the bounding box by calculating the Bhattacharyya distance between the pre-computed mean distribution of instances and the instance distribution predicted by the model. Experimental results show that the proposed method can effectively reduce the overlapping bounding boxes and retain bounding box annotations that are closer to the ground truth. Based on artefact detection, under the guidance of professional endoscopists, a quality score assessment is developed to calculate the quality score of the entire image by setting different weights for different regions, areas, and artefact types.
- **Chapter 4** of this thesis first presents an end-to-end ulcerative colitis classification method utilizing supervised learning. The proposed method employs additive angular margin loss to enhance inter-class separation, significantly boosting the classification performance. Due to data limitations, a novel SSL-CPCD method is developed for learning semantically meaningful features from unlabeled data, improving the performance of subsequent downstream tasks such as ulcerative colitis classification. The proposed SSL-CPCD method integrates a composite loss function to minimise the noise contrast estimation loss across the target image-level and patch-level embedding while incorporating an extra angular margin to the cosine similarity to intensify the

penalization of the decision boundary. Comprehensive comparisons with state-of-the-art methods and extended out-of-sample studies on multiple different datasets demonstrate the superiority of the proposed method.

- **Chapter 5** of this thesis first presents the performance of SSL-CPCD method in polyp detection and segmentation. For the challenging problem of video polyp segmentation, a novel end-to-end framework, SSTFB, utilizing self-supervised and normalised self-attention blocks. These blocks are optimally configured and incorporate feature branching to meet real-time performance criteria. The self-supervised learning scheme serves as an auxiliary task within end-to-end framework, enhancing the feature representation. Furthermore, the normalised self-attention module employs feature aggregation and correlation measurement to effectively harness the spatiotemporal relationships of features, thereby enabling the learning of more meaningful features through feature branching and concatenation techniques.

## 6.2 Limitations and directions for future work

Apart from the core contributions presented in this thesis, several other key tasks in endoscopic vision remain under-explored and are addressed in the following sections. These include enhancing image quality through artefact restoration, improving UC grading by adopting more fine-grained clinical scoring systems, and quantifying disease burden via 3D reconstruction of the intestinal mucosa. Each of these directions is essential to further improve the clinical utility and robustness of AI-assisted endoscopic systems, and represents a logical extension of our current work.

### 6.2.1 Image quality enhancement based on artefact restoration

Chapter 3 presented an image quality assessment framework based on automatic artefact detection. For video frames with low quality scores, restoring the detected artefacts using image restoration techniques can effectively avoid artefacts from obscuring the underlying tissue, better display the mucosal structure, lesions and

other important features. This enhancement potentially reduces misdiagnosis rates, even speeds up the diagnosis and shortens the duration of endoscopic examinations. Furthermore, image quality plays a vital role in endoscopic surgery. For example, in the case of polypectomy or biopsy, image restoration helps provide a clearer view, helping endoscopists to accurately navigate and operate instruments, and make more accurate real-time decisions, which is crucial for patient safety and successful surgical outcomes.

Different image restoration strategies should be adopted for different types of artefacts. For specular reflections and artefacts, which present as pixel missing issues, Conditional Generative Adversarial Nets (CGAN) with  $l1$  context loss and generation loss can avoid the discontinuity retention term to prevent over-smoothing. Saturation and low contrast problems are typically caused by illumination changes and occupy a large image area. Substituting the  $l1$  context loss in CGAN with  $l2$  provides more effectively capture the deviation between normal lighting conditions and saturated and low-contrast conditions. For motion blur, Generative LatEnt bANk (GLEAN) based on StyleGAN, pre-trained and can efficiently restore high-resolution clear images.

### 6.2.2 More comprehensive grading strategy for ulcerative colitis

Chapter 4 presented the comprehensive study in ulcerative colitis grading task utilizing the MES. In consultation with a dedicated endoscopist (Prof. Simon Leedham), MES is often used for the initial evaluation of UC, while a more detailed evaluation uses the Ulcerative Colitis Endoscopic Index of Severity (UCEIS). UCEIS is designed to provide a more standardised and detailed assessment compared to MES. UCEIS includes three specific variables: vascular pattern, bleeding, and erosions and ulcers, each graded on a clearly defined scale. This granularity helps in reducing inter-observer variability and increases the specificity of the assessment. UCEIS has been shown to be highly sensitive to changes in the mucosal appearance, making it particularly useful in clinical trials and for monitoring disease progression or response

to treatment. MES divides UC severity into three levels, and UCEIS further refines the classification and divides UC severity into 4 strata and 9 levels: remission (UCEIS 0–1); mild (UCEIS 2–4); moderate (UCEIS 5–6); and severe (UCEIS 7–8).

However, there is a lack of sufficient datasets for deep learning training, and obtaining ground truth annotations is an expensive process. Therefore, it is necessary to build a UC dataset scored by UCEIS. After obtaining the dataset, we will test our proposed method and develop a more robust multi-classification algorithm.

### 6.2.3 Quantification of disease burden

The intestinal mucosa is a very important dense connective tissue, which maintains the homeostasis in the intestine together with the intestinal flora. Mucosa is easily affected by diseases, such as inflammation (ulcerative colitis, Crohn’s disease, etc.) and intestinal mucosal damage can lead to the formation of polyps on the mucosal surface. In colonoscopy, it is very important to identify the intestinal mucosa. The reason is that abnormalities such as polyps and adenomas in the colon and rectum can often be identified on the mucosal surface. Therefore, it is necessary to quantify the intestinal mucosal surface in the lesion area. Most current literature focus on 3D reconstruction of the colonic surface. Ma et al. [195] developed a deep learning-driven dense SLAM (simultaneous localization and mapping) system that can generate camera trajectories and dense reconstructed surfaces for the colon, and uses recurrent neural networks (RNN) to predict consistent depth maps and camera pose for consecutive frames. Other methods that use depth estimation showing successful 3D reconstruction of GI organs (e.g. Mahmood’s work [196] and Ali’s work [197] have already used depth estimation oesophagus).

Developed method is used to collate images using pixel level registration and depth is used to project to the 3D world co-ordinate system. Endoscope are closed to scene which can be considered planar or quasi-planar. If away and deformation as in real scenario then the scene becomes non-planar and that will require pixel-to-pixel registration to achieve accurate 3D mosaics. Here the assumption is that the scene is planar. However, the colonic surface is non-planar scene with large deformations.



So, the idea will be to use pixel-level registration for our work. We plan to build a deep learning-based framework as a depth and pose estimation network. The generated predicted depth map and the camera pose for each image in the video, which can show the distance between the target area and the camera. Combined with the deformable registration network, the non-planar and complex colonic surface will be 3D quantification. After quantifying the intestinal mucosa through the registration network, we plan to train a region proposal network to count the polyps in the intestine. The region proposal network can be used on the feature map extracted by the depth estimation network to further locate the polyp region.

### 6.3 Conclusion

Computer-aided endoscopy is expected to have a transformative impact on clinical practice by enhancing diagnostic accuracy, improving procedural efficiency, and potentially reducing human error. However, before this technology can be fully integrated into everyday clinical use, several critical challenges must be addressed. This thesis comprehensively explores several challenges faced during endoscopy, and the proposed image-driven method based on deep learning can effectively address these challenges. Image quality assessment is an important prerequisite for endoscopy. Higher image quality can not only reduce the visual impact of artefacts on endoscopists, but also avoid multiple follow-up visits for patients. Automatic detection of lesions such as IBD can assist and prompt endoscopists to pay attention to areas that may be overlooked and reduce the rate of missed disease detection. In particular, the automatic and accurate segmentation of polyps can provide important parameters for subsequent polypectomy. In addition to exploring the details of image analysis technology and building a more comprehensive endoscopic dataset, future work will also aim to further examine the performance of the proposed method in clinical practice, thereby accelerating the application of this technology in clinical endoscopy. Finally, based on the research in this thesis, I would like to emphasise that the development of image-driven medical image

analysis methods can greatly reduce the workload of endoscopists and improve diagnosis and subsequent treatment.

## References

- [1] Hamid Asadzadeh Aghdaei et al. “Polyp detection rate and pathological features in patients undergoing a comprehensive colonoscopy screening”. In: *World journal of gastrointestinal pathophysiology* 8.1 (2017), p. 3.
- [2] Unnati Bhalerao et al. “Diagnosis of colorectal cancer using molecular techniques”. In: *Colon Cancer Diagnosis and Therapy: Volume 2*. Springer, 2021, pp. 143–170.
- [3] Hanna Borgli et al. “HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy”. In: *Scientific data* 7.1 (2020), pp. 1–14.
- [4] Azadeh Khanicheh and Amandeep K Shergill. “Endoscope design for the future”. In: *Techniques in gastrointestinal endoscopy* 21.3 (2019), pp. 167–173.
- [5] Rudolph Schindler and George B Eusterman. “Gastrosocopy: the endoscopic study of gastric pathology”. In: *Annals of Surgery* 106.5 (1937), p. 958.
- [6] Douglas K Rex et al. “Quality indicators for colonoscopy”. In: *Gastrointestinal endoscopy* 63.4 (2006), S16–S28.
- [7] W Sircus. “Milestones in the evolution of endoscopy: a short history.” In: *The journal of the Royal College of Physicians of Edinburgh* 33.2 (2003), pp. 124–134.
- [8] Gavriel Iddan et al. “Wireless capsule endoscopy”. In: *Nature* 405.6785 (2000), pp. 417–417.
- [9] Albert J de Groof et al. “Deep-learning system detects neoplasia in patients with Barrett’s esophagus with higher accuracy than endoscopists in a multistep training and validation study with benchmarking”. In: *Gastroenterology* 158.4 (2020), pp. 915–929.
- [10] Yuan-Yen Chang et al. “Development and validation of a deep learning-based algorithm for colonoscopy quality assessment”. In: *Surgical Endoscopy* 36.9 (2022), pp. 6446–6455.
- [11] Ziang Xu et al. “Self-supervised approach for a fully assistive esophageal surveillance: Quality, anatomy and neoplasia guidance”. In: *MICCAI Workshop on Cancer Prevention through Early Detection*. Springer. 2022, pp. 14–23.
- [12] Andreas Sturm et al. “ECCO-ESGAR Guideline for Diagnostic Assessment in IBD Part 2: IBD scores and general principles and technical aspects”. In: *Journal of Crohn’s and Colitis* 13.3 (2019), pp. 273–284.
- [13] Arie Levine et al. “Pediatric modification of the Montreal classification for inflammatory bowel disease: the Paris classification”. In: *Inflammatory bowel diseases* 17.6 (2011), pp. 1314–1321.
- [14] Elizabeth R Paine. “Colonoscopic evaluation in ulcerative colitis”. In: *Gastroenterology report* 2.3 (2014), pp. 161–168.

- [15] Philip H Gordon and Santhat Nivatvongs. *Principles and practice of surgery for the colon, rectum, and anus*. crc press, 2007.
- [16] Joel S Levine and Dennis J Ahnen. “Adenomatous polyps of the colon”. In: *New England Journal of Medicine* 355.24 (2006), pp. 2551–2557.
- [17] Matti Pietikäinen, Topi Mäenpää, and Jaakko Viertola. “Color texture classification with color histograms and local binary patterns”. In: *Workshop on texture analysis in machine vision*. Vol. 1. Citeseer. 2002, pp. 109–112.
- [18] Aditya Vailaya et al. “Image classification for content-based indexing”. In: *IEEE transactions on image processing* 10.1 (2001), pp. 117–130.
- [19] Annamalai Suresh, KL Shunmuganathan, et al. “Image texture classification using gray level co-occurrence matrix based statistical features”. In: *European Journal of Scientific Research* 75.4 (2012), pp. 591–597.
- [20] Zhenhua Guo, Lei Zhang, and David Zhang. “A completed modeling of local binary pattern operator for texture classification”. In: *IEEE transactions on image processing* 19.6 (2010), pp. 1657–1663.
- [21] Cecilia R Aragon and David Bradburn Aragon. “A fast contour descriptor algorithm for supernova image classification”. In: *Real-Time Image Processing 2007*. Vol. 6496. SPIE. 2007, pp. 61–72.
- [22] VF Leavers. “Which hough transform?” In: *CVGIP: Image understanding* 58.2 (1993), pp. 250–264.
- [23] Hervé Abdi and Lynne J Williams. “Principal component analysis”. In: *Wiley interdisciplinary reviews: computational statistics* 2.4 (2010), pp. 433–459.
- [24] Suresh Balakrishnama and Aravind Ganapathiraju. “Linear discriminant analysis-a brief tutorial”. In: *Institute for Signal and information Processing* 18.1998 (1998), pp. 1–8.
- [25] Marti A. Hearst et al. “Support vector machines”. In: *IEEE Intelligent Systems and their applications* 13.4 (1998), pp. 18–28.
- [26] Leif E Peterson. “K-nearest neighbor”. In: *Scholarpedia* 4.2 (2009), p. 1883.
- [27] GM Cramer, RA Ford, and RL Hall. “Estimation of toxic hazard—a decision tree approach”. In: *Food and cosmetics toxicology* 16.3 (1976), pp. 255–276.
- [28] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32.
- [29] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.
- [30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012).
- [31] Keiron O’shea and Ryan Nash. “An introduction to convolutional neural networks”. In: *arXiv preprint arXiv:1511.08458* (2015).
- [32] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

- [33] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).
- [34] Christian Szegedy et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.
- [35] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [36] Gao Huang et al. “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.
- [37] Mingxing Tan and Quoc Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114.
- [38] Teng Zhou et al. “Quantitative analysis of patients with celiac disease by video capsule endoscopy: A deep learning method”. In: *Computers in biology and medicine* 85 (2017), pp. 1–6.
- [39] Sungheon Park, Myunggi Lee, and Nojun Kwak. “Polyp detection in colonoscopy videos using deeply-learned hierarchical features”. In: *Seoul National University* (2015).
- [40] Eduardo Ribeiro et al. “Exploring deep learning and transfer learning for colonic polyp classification”. In: *Computational and mathematical methods in medicine* 2016 (2016).
- [41] Lequan Yu et al. “Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos”. In: *IEEE journal of biomedical and health informatics* 21.1 (2016), pp. 65–75.
- [42] Sinno Jialin Pan and Qiang Yang. “A survey on transfer learning”. In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.
- [43] Ruikai Zhang et al. “Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain”. In: *IEEE journal of biomedical and health informatics* 21.1 (2016), pp. 41–47.
- [44] Rajesh Godasu, David Zeng, and Kruttika Suttrave. “Transfer learning in medical image classification: Challenges and opportunities”. In: *Transfer* 5 (2020), pp. 28–2020.
- [45] Michael F Byrne et al. “Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model”. In: *Gut* 68.1 (2019), pp. 94–100.
- [46] Albert J. de Groof et al. “Deep-Learning System Detects Neoplasia in Patients With Barrett’s Esophagus With Higher Accuracy Than Endoscopists in a Multistep Training and Validation Study With Benchmarking”. In: *Gastroenterology* 158.4 (2020), 915–929.e4.
- [47] Gaoshuang Liu et al. “Automatic classification of esophageal lesions in endoscopic images using a convolutional neural network”. In: *Annals of Translational Medicine* 8.7 (2020).

- [48] Alanna Ebigbo et al. “Computer-aided diagnosis using deep learning in the evaluation of early oesophageal adenocarcinoma”. In: *Gut* 68.7 (2019), pp. 1143–1145.
- [49] Zhan Wu et al. “ELNet: Automatic classification and segmentation for esophageal lesions using convolutional neural network”. In: *Medical Image Analysis* 67 (2021), p. 101838.
- [50] Mohamed Hussein et al. “Deep Neural Network For The Detection Of Early Neoplasia In Barrett’s Oesophagus”. In: *Gastrointestinal Endoscopy* 91.6 (2020), AB250.
- [51] Qi He et al. “Deep learning-based anatomical site classification for upper gastrointestinal endoscopy”. In: *International journal of computer assisted radiology and surgery* 15.7 (2020), pp. 1085–1094.
- [52] Park Jung-Whan et al. “Automatic Anatomical Classification Model of Esophagogastroduodenoscopy Images Using Deep Convolutional Neural Networks for Guiding Endoscopic Photodocumentation”. In: *Journal of The Korea Society of Computer and Information* 26.3 (2021), pp. 19–28.
- [53] Hirotoshi Takiyama et al. “Automatic anatomical classification of esophagogastroduodenoscopy images using deep convolutional neural networks”. In: *Scientific reports* 8.1 (2018), pp. 1–8.
- [54] Tsuyoshi Ozawa et al. “Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis”. In: *Gastrointestinal endoscopy* 89.2 (2019), pp. 416–421.
- [55] Ryan W Stidham et al. “Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis”. In: *JAMA network open* 2.5 (2019), e193963–e193963.
- [56] Ali Alammari et al. “Classification of ulcerative colitis severity in colonoscopy videos using CNN”. In: *Proceedings of the 9th international conference on information management and engineering*. 2017, pp. 139–144.
- [57] Benjamin Gutierrez Becker et al. “Training and deploying a deep learning model for endoscopic severity grading in ulcerative colitis using multicenter clinical trial data”. In: *Therapeutic advances in gastrointestinal endoscopy* 14 (2021), p. 2631774521990623.
- [58] Paul Viola and Michael Jones. “Rapid object detection using a boosted cascade of simple features”. In: *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*. Vol. 1. Ieee. 2001, pp. I–I.
- [59] Navneet Dalal and Bill Triggs. “Histograms of oriented gradients for human detection”. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*. Vol. 1. Ieee. 2005, pp. 886–893.
- [60] William T Freeman and Michal Roth. “Orientation histograms for hand gesture recognition”. In: *International workshop on automatic face and gesture recognition*. Vol. 12. Citeseer. 1995, pp. 296–301.
- [61] Junjie Yan et al. “The fastest deformable part model for object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 2497–2504.

- [62] Joseph Redmon et al. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.
- [63] Wei Liu et al. “Ssd: Single shot multibox detector”. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer. 2016, pp. 21–37.
- [64] M. Everingham et al. *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results*. <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>.
- [65] Tsung-Yi Lin et al. “Focal loss for dense object detection”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.
- [66] Mingxing Tan, Ruoming Pang, and Quoc V Le. “Efficientdet: Scalable and efficient object detection”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10781–10790.
- [67] Ross Girshick et al. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.
- [68] Ross Girshick. “Fast r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448.
- [69] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *arXiv preprint arXiv:1506.01497* (2015).
- [70] Sharib Ali et al. “An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy”. In: *Scientific reports* 10.1 (2020), pp. 1–15.
- [71] Gorkem Polat et al. “Endoscopic Artefact Detection with Ensemble of Deep Neural Networks and False Positive Elimination.” In: *EndoCV@ ISBI*. 2020, pp. 8–12.
- [72] Pengyi Zhang, Xiaoqiong Li, and Y Zhong. “Ensemble mask-aided r-cnn”. In: *ISBI*. 2018, pp. 6154–6162.
- [73] Zhaowei Cai and Nuno Vasconcelos. “Cascade r-cnn: Delving into high quality object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 6154–6162.
- [74] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer. 2014, pp. 740–755.
- [75] Hei Law and Jia Deng. “Cornersnet: Detecting objects as paired keypoints”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 734–750.
- [76] Kaiwen Duan et al. “Centernet: Keypoint triplets for object detection”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6569–6578.
- [77] Anand Subramanian and Koushik Srivatsan. “Exploring Deep Learning Based Approaches for Endoscopic Artefact Detection and Segmentation.” In: *EndoCV@ ISBI*. 2020, pp. 51–56.

- [78] Ilkay Oksuz et al. “Artefact detection in video endoscopy using retinanet and focal loss function”. In: *Proceedings of the 2019 Challenge on Endoscopy Artefacts Detection (EAD2019), Venice, Italy, 8th April* 2366 (2019).
- [79] Mohammad Azam Khan and Jaegul Choo. “Multi-class artefact detection in video endoscopy via convolution neural networks”. In: *Proceedings of the 1st International Workshop and Challenge on Computer Vision in Endoscopy*. 2019.
- [80] Suhui Yang and Guanju Cheng. “Endoscopic artefact detection and segmentation with deep convolutional neural network”. In: *Proceedings of the 1st International Workshop and Challenge on Computer Vision in Endoscopy*. 2019.
- [81] Xiaokang Wang and Chunqing Wang. “Detect artefacts of various sizes on the right scale for each class in video endoscopy”. In: *Proceedings of the 2019 Challenge on Endoscopy Artefacts Detection (EAD2019), Venice, Italy, 8th April* 2366 (2019).
- [82] Seiryō Watanabe, Shigeto Seno, and Hideo Matsuda. “DNN Models and Postprocessing Thresholds for Endoscopy Artifact Detection in Practice”. In: *EAD2019-CEUR Workshop Proceedings*. Vol. 2366. 2019.
- [83] Zhaoyu Yuan et al. “EAD-YOLO: Improved YOLOv5 for Endoscopic Artefact Detection”. In: *2023 International Conference on Communications, Computing and Artificial Intelligence (CCCAI)*. IEEE. 2023, pp. 151–158.
- [84] Ji Young Lee et al. “Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets”. In: *Scientific reports* 10.1 (2020), pp. 1–9.
- [85] Xu Zhang et al. “Real-time gastric polyp detection using convolutional neural networks”. In: *PloS one* 14.3 (2019), e0214133.
- [86] Hemin Ali Qadir et al. “Polyp detection and segmentation using mask R-CNN: Does a deeper feature extractor CNN always perform better?” In: *2019 13th International Symposium on Medical Information and Communication Technology (ISMICT)*. 2019, pp. 1–6.
- [87] Younghak Shin et al. “Automatic colon polyp detection using region based deep CNN and post learning approaches”. In: *IEEE Access* 6 (2018), pp. 40950–40962.
- [88] Xinyu Liu et al. “Consolidated domain adaptive detection and localization framework for cross-device colonoscopic images”. In: *Medical image analysis* 71 (2021), p. 102052.
- [89] Adrian Krenzer et al. “A real-time polyp-detection system with clinical application in colonoscopy using deep convolutional neural networks”. In: *Journal of imaging* 9.2 (2023), p. 26.
- [90] Ralf Kohler. “A segmentation system based on thresholding”. In: *Computer Graphics and Image Processing* 15.4 (1981), pp. 319–338.
- [91] Aly A Farag. “Edge-based image segmentation”. In: *Remote sensing reviews* 6.1 (1992), pp. 95–121.
- [92] S Karthick, K Sathiyasekar, and A Puraneeswari. “A survey based on region based segmentation”. In: *International Journal of Engineering Trends and Technology* 7.3 (2014), pp. 143–147.



- [93] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. “Snakes: Active contour models”. In: *International journal of computer vision* 1.4 (1988), pp. 321–331.
- [94] Stanley Osher and Ronald P Fedkiw. “Level set methods: an overview and some recent results”. In: *Journal of Computational physics* 169.2 (2001), pp. 463–502.
- [95] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.
- [96] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-net: Convolutional networks for biomedical image segmentation”. In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III* 18. Springer. 2015, pp. 234–241.
- [97] Zongwei Zhou et al. “Unet++: A nested u-net architecture for medical image segmentation”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings* 4. Springer. 2018, pp. 3–11.
- [98] Debesh Jha et al. “Resunet++: An advanced architecture for medical image segmentation”. In: *2019 IEEE International Symposium on Multimedia (ISM)*. IEEE. 2019, pp. 225–2255.
- [99] Chang Wang et al. “Dense U-net based on patch-based learning for retinal vessel segmentation”. In: *Entropy* 21.2 (2019), p. 168.
- [100] Alexander V Mamonov et al. “Automated polyp detection in colon capsule endoscopy”. In: *IEEE transactions on medical imaging* 33.7 (2014), pp. 1488–1502.
- [101] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. “Automated polyp detection in colonoscopy videos using shape and context information”. In: *IEEE transactions on medical imaging* 35.2 (2015), pp. 630–644.
- [102] Ruikai Zhang et al. “Polyp detection during colonoscopy using a regression-based convolutional neural network with a tracker”. In: *Pattern recognition* 83 (2018), pp. 209–219.
- [103] Balamurali Murugesan et al. “Psi-Net: Shape and boundary aware joint multi-task deep network for medical image segmentation”. In: *2019 41st Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE. 2019, pp. 7223–7226.
- [104] Yuqi Fang et al. “Selective feature aggregation network with area-boundary constraints for polyp segmentation”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I* 22. Springer. 2019, pp. 302–310.
- [105] Deng-Ping Fan et al. “Pranet: Parallel reverse attention network for polyp segmentation”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI* 23. Springer. 2020, pp. 263–273.

- [106] Bo Dong et al. “Polyp-pvt: Polyp segmentation with pyramid vision transformers”. In: *arXiv preprint arXiv:2108.06932* (2021).
- [107] Md Mostafijur Rahman and Radu Marculescu. “Medical image segmentation via cascaded attention decoding”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 6222–6231.
- [108] Bin Xiao et al. “CTNet: Contrastive Transformer Network for Polyp Segmentation”. In: *IEEE Transactions on Cybernetics* (2024).
- [109] Na Ta et al. “BLE-Net: boundary learning and enhancement network for polyp segmentation”. In: *Multimedia Systems* 29.5 (2023), pp. 3041–3054.
- [110] Zihuan Qiu et al. “BDG-Net: boundary distribution guided network for accurate polyp segmentation”. In: *Medical Imaging 2022: Image Processing*. Vol. 12032. SPIE. 2022, pp. 792–799.
- [111] Lu Lu et al. “A multi-scale perceptual polyp segmentation network based on boundary guidance”. In: *Image and Vision Computing* 138 (2023), p. 104811.
- [112] Samir Jain et al. “CoInNet: A Convolution-Involution Network with a Novel Statistical Attention for Automatic Polyp Segmentation”. In: *IEEE Transactions on Medical Imaging* (2023).
- [113] Tao Zhou et al. “Cross-level feature aggregation network for polyp segmentation”. In: *Pattern Recognition* 140 (2023), p. 109555.
- [114] Xinkai Zhao et al. “Semi-supervised spatial temporal attention network for video polyp segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 456–466.
- [115] Xiaolong Wang et al. “Non-local neural networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7794–7803.
- [116] Ge-Peng Ji et al. “Progressively normalized self-attention network for video polyp segmentation”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer. 2021, pp. 142–152.
- [117] Ge-Peng Ji et al. “Video polyp segmentation: A deep learning perspective”. In: *Machine Intelligence Research* (2022), pp. 1–19.
- [118] Qi Chang et al. “ESFPNet: efficient deep learning architecture for real-time lesion segmentation in autofluorescence bronchoscopic video”. In: *Medical Imaging 2023: Biomedical Applications in Molecular, Structural, and Functional Imaging*. Vol. 12468. SPIE. 2023, p. 1246803.
- [119] Yefan Xiao et al. “Icbnet: Iterative context-boundary feedback network for polyp segmentation”. In: *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2022, pp. 1297–1304.
- [120] Ishan Misra and Laurens van der Maaten. “Self-supervised learning of pretext-invariant representations”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 6707–6717.
- [121] Ting Chen et al. “A simple framework for contrastive learning of visual representations”. In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

- [122] Kaiming He et al. “Momentum contrast for unsupervised visual representation learning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738.
- [123] Jane Bromley et al. “Signature verification using a " siamese" time delay neural network”. In: *Advances in neural information processing systems* 6 (1993).
- [124] Sumit Chopra, Raia Hadsell, and Yann LeCun. “Learning a similarity metric discriminatively, with application to face verification”. In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*. Vol. 1. IEEE. 2005, pp. 539–546.
- [125] Raia Hadsell, Sumit Chopra, and Yann LeCun. “Dimensionality reduction by learning an invariant mapping”. In: *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR’06)*. Vol. 2. IEEE. 2006, pp. 1735–1742.
- [126] Kaiming He et al. “Masked autoencoders are scalable vision learners”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 16000–16009.
- [127] Mehdi Noroozi and Paolo Favaro. “Unsupervised learning of visual representations by solving jigsaw puzzles”. In: *European conference on computer vision*. Springer. 2016, pp. 69–84.
- [128] Xudong Wang et al. “Unsupervised feature learning by cross-level instance-group discrimination”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 12586–12595.
- [129] Jean-Bastien Grill et al. “Bootstrap your own latent-a new approach to self-supervised learning”. In: *Advances in neural information processing systems* 33 (2020), pp. 21271–21284.
- [130] Mathilde Caron et al. “Emerging properties in self-supervised vision transformers”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–9660.
- [131] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).
- [132] Hangbo Bao et al. “Beit: Bert pre-training of image transformers”. In: *arXiv preprint arXiv:2106.08254* (2021).
- [133] Zhenda Xie et al. “Simmim: A simple framework for masked image modeling”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 9653–9663.
- [134] Jihao Liu et al. “MixMAE: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 6252–6261.
- [135] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. “Spatial transformer networks”. In: *Advances in neural information processing systems* 28 (2015).
- [136] Jie Hu, Li Shen, and Gang Sun. “Squeeze-and-excitation networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7132–7141.

- [137] Sanghyun Woo et al. “Cbam: Convolutional block attention module”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.
- [138] Tal Ridnik et al. “Imagenet-21k pretraining for the masses”. In: *arXiv preprint arXiv:2104.10972* (2021).
- [139] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).
- [140] Sharib Ali et al. “Deep learning for detection and segmentation of artefact and disease instances in gastrointestinal endoscopy”. In: *Medical Image Analysis* (2021), p. 102002.
- [141] F Javier Sánchez et al. “Bright spot regions segmentation and classification for specular highlights detection in colonoscopy videos”. In: *Machine Vision and Applications* 28.8 (2017), pp. 917–936.
- [142] Shu Liu et al. “Path aggregation network for instance segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 8759–8768.
- [143] Alexander Neubeck and Luc Van Gool. “Efficient non-maximum suppression”. In: *18th International Conference on Pattern Recognition (ICPR’06)*. Vol. 3. IEEE. 2006, pp. 850–855.
- [144] Thomas Kailath. “The divergence and Bhattacharyya distance measures in signal selection”. In: *IEEE transactions on communication technology* 15.1 (1967), pp. 52–60.
- [145] Guy Barrett Coleman and Harry C Andrews. “Image segmentation by clustering”. In: *Proceedings of the IEEE* 67.5 (1979), pp. 773–785.
- [146] Joseph Redmon and Ali Farhadi. “Yolov3: An incremental improvement”. In: *arXiv preprint arXiv:1804.02767* (2018).
- [147] G Polat et al. *Labeled Images for Ulcerative Colitis (LIMUC) Dataset*. 2022.
- [148] Geert D’haens et al. “A review of activity indices and efficacy end points for clinical trials of medical therapy in adults with ulcerative colitis”. In: *Gastroenterology* 132.2 (2007), pp. 763–786.
- [149] Jiankang Deng et al. “Arcface: Additive angular margin loss for deep face recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4690–4699.
- [150] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.
- [151] Alissa Walsh, Rebecca Palmer, and Simon Travis. “Mucosal healing as a target of therapy for colonic inflammatory bowel disease and methods to score disease activity”. In: *Gastrointestinal Endoscopy Clinics* 24.3 (2014), pp. 367–378.
- [152] Mark Sandler et al. “Mobilenetv2: Inverted residuals and linear bottlenecks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.

- [153] Chaitanya Kaul et al. “FocusNet: An attention-based fully convolutional network for medical image segmentation”. In: *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*. 2019, pp. 455–458.
- [154] Shekoofeh Azizi et al. “Big Self-Supervised Models Advance Medical Image Classification”. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 3458–3468.
- [155] Liang Chen et al. “Self-supervised learning for medical image analysis using image context restoration”. In: *Medical image analysis* 58 (2019), p. 101539.
- [156] Sharib Ali. “Where do we stand in AI for endoscopic image analysis? Deciphering gaps and future directions”. In: *npj Digital Medicine* 5.1 (Dec. 2022), p. 184.
- [157] Hong Liu et al. “Self-supervised Learning is More Robust to Dataset Imbalance”. In: *International Conference on Learning Representations*. 2022.
- [158] Mathias Lux and Michael Riegler. “Annotation of endoscopic videos on mobile devices: a bottom-up approach”. In: *Proceedings of the 4th ACM Multimedia Systems Conference*. 2013, pp. 141–145.
- [159] Hong-Yu Zhou et al. “Preservational learning improves self-supervised medical image models by reconstructing diverse contexts”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 3499–3509.
- [160] Patrick Dermeyer, Angad Kalra, and Matt Schwartz. “EndoDINO: A Foundation Model for GI Endoscopy”. In: *arXiv preprint arXiv:2501.05488* (2025).
- [161] Zhao Wang et al. “Foundation model for endoscopy video analysis via large-scale self-supervised pre-train”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 101–111.
- [162] Weiran Huang et al. “Towards the Generalization of Contrastive Self-Supervised Learning”. In: *The Eleventh International Conference on Learning Representations (ICLR)*. 2023.
- [163] Xinrui Zhuang et al. “Self-supervised feature learning for 3d medical images by playing a rubik’s cube”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 420–428.
- [164] Xuan-Bac Nguyen et al. “Self-supervised learning based on spatial awareness for medical image analysis”. In: *IEEE Access* 8 (2020), pp. 162973–162981.
- [165] Zeng Zeng et al. “Sese-net: Self-supervised deep learning for segmentation”. In: *Pattern Recognition Letters* 128 (2019), pp. 23–29.
- [166] Ozan Ciga et al. “Self supervised contrastive learning for digital histopathology”. In: *Machine Learning with Applications* 7 (2022), p. 100198.
- [167] Liang Chen et al. “Self-supervised learning for medical image analysis using image context restoration”. In: *Medical Image Analysis* 58 (2019), p. 101539. URL: <https://www.sciencedirect.com/science/article/pii/S1361841518304699>.
- [168] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605.
- [169] Zhirong Wu et al. “Unsupervised Feature Learning via Non-Parametric Instance Discrimination”. In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 2018, pp. 3733–3742.

- [170] Ziang Xu et al. “Patch-level instance-group discrimination with pretext-invariant learning for colitis scoring”. In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2022, pp. 101–110.
- [171] Jiankang Deng et al. “ArcFace: Additive Angular Margin Loss for Deep Face Recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.
- [172] Debesh Jha et al. “Kvasir-SEG: A segmented polyp dataset”. In: *International Conference on Multimedia Modeling*. 2020, pp. 451–462.
- [173] Jorge Bernal et al. “WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians”. In: *Computerized medical imaging and graphics* 43 (2015), pp. 99–111.
- [174] Michael Gutmann and Aapo Hyvärinen. “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models”. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 297–304.
- [175] Michael Gutmann and Aapo Hyvärinen. “Noise-contrastive estimation: A new estimation principle for unnormalized statistical models”. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Vol. 9. 2010, pp. 297–304.
- [176] Adam Paszke et al. “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems* 32 (2019).
- [177] Ching-Yao Chuang et al. “Debiased contrastive learning”. In: *Advances in neural information processing systems* 33 (2020), pp. 8765–8775.
- [178] Mingxing Tan and Quoc Le. “Efficientnetv2: Smaller models and faster training”. In: *International conference on machine learning*. PMLR. 2021, pp. 10096–10106.
- [179] Mingxing Tan and Quoc Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Sept. 2019, pp. 6105–6114.
- [180] Jie Hu, Li Shen, and Gang Sun. “Squeeze-and-Excitation Networks”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7132–7141.
- [181] Fons van der Sommen et al. “Machine learning in GI endoscopy: practical guidance in how to interpret a novel field”. en. In: *Gut* 69.11 (Nov. 2020), pp. 2035–2045.
- [182] Sharib Ali et al. “Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge”. In: *Sci Rep* 14 (2024), p. 2032.
- [183] Masashi Misawa et al. “Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video)”. In: *Gastrointestinal endoscopy* 93.4 (2021), pp. 960–967.

- [184] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. “Road extraction by deep residual u-net”. In: *IEEE Geoscience and Remote Sensing Letters* 15.5 (2018), pp. 749–753.
- [185] Juan Silva et al. “Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer”. In: *International journal of computer assisted radiology and surgery* 9 (2014), pp. 283–293.
- [186] Fatima A Hagggar and Robin P Boushey. “Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors”. In: *Clinics in colon and rectal surgery* 22.04 (2009), pp. 191–197.
- [187] Ashish Jaiswal et al. “A survey on contrastive self-supervised learning”. In: *Technologies* 9.1 (2020), p. 2.
- [188] Shang-Hua Gao et al. “Res2net: A new multi-scale backbone architecture”. In: *IEEE transactions on pattern analysis and machine intelligence* 43.2 (2019), pp. 652–662.
- [189] Jun Wei et al. “Shallow attention network for polyp segmentation”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*. Springer. 2021, pp. 699–708.
- [190] Nikhil Kumar Tomar et al. “TGANet: text-guided attention for improved polyp segmentation”. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*. Springer. 2022, pp. 151–160.
- [191] Deng-Ping Fan et al. “Structure-measure: A new way to evaluate foreground maps”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 4548–4557.
- [192] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. “How to evaluate foreground maps?” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 248–255.
- [193] Radhakrishna Achanta et al. “Frequency-tuned salient region detection”. In: *2009 IEEE conference on computer vision and pattern recognition*. IEEE. 2009, pp. 1597–1604.
- [194] Deng-Ping Fan et al. “Enhanced-alignment measure for binary foreground map evaluation”. In: *arXiv preprint arXiv:1805.10421* (2018).
- [195] Ruibin Ma et al. “Real-time 3D reconstruction of colonoscopic surfaces for determining missing regions”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 573–582.
- [196] Faisal Mahmood and Nicholas J Durr. “Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy”. In: *Medical image analysis* 48 (2018), pp. 230–243.
- [197] Sharib Ali et al. “A Pilot Study on Automatic Three-Dimensional Quantification of Barrett’s Esophagus for Risk Stratification and Therapy Monitoring”. In: *Gastroenterology* 161.3 (2021), 865–878.e8.