White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Leeds.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/213496/

Version: Accepted Version

Article:
Xu, Z., Rittscher, J. and Ali, S. (2024) SSL-CPCD: Self-supervised learning with composite pretext-class discrimination for improved generalisability in endoscopic image analysis. IEEE Transactions on Medical Imaging. ISSN: 0278-0062

https://doi.org/10.1109/tmi.2024.3411933

© 2024 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



SSL-CPCD: Self-supervised learning with composite pretext-class discrimination for improved generalisability in endoscopic image analysis

Ziang Xu, Jens Rittscher, and Sharib Ali

***Abstract*— Data-driven methods have shown tremendous progress in medical image analysis. In this context, deep learning-based supervised methods are widely popular. However, they require a large amount of training data and face issues in generalisability to unseen datasets that hinder clinical translation. Endoscopic imaging data is characterised by large inter- and intra-patient variability that makes these models more challenging to learn representative features for downstream tasks. Thus, despite the publicly available datasets and datasets that can be generated within hospitals, most supervised models still underperform. While self-supervised learning has addressed this problem to some extent in natural scene data, there is a considerable performance gap in the medical image domain. In this paper, 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. Our results demonstrate significant improvement on all metrics over the state-of-the-art (SOTA) methods on the test set from the same and diverse datasets. We evaluated our approach for classification, detection, and segmentation. SSL-CPCD attains notable Top 1 accuracy of 79.77% in ulcerative colitis classification, an 88.62% mean average precision (mAP) for detection, and an 82.32% dice similarity coefficient for segmentation tasks. These represent improvements of over 4%, 2%, and 3%, respectively, compared to the baseline architectures. We demonstrate that our method generalises better than all SOTA methods to unseen datasets, reporting over 7% improvement.**

***Index Terms*— Deep learning, contrastive loss, endoscopy data, generalisation, self-supervised learning**

Manuscript received 20 February 2024; accepted 5 June 2024. (Corresponding author: Sharib Ali.)

The research was supported by the National Institute for Health Research Oxford Biomedical Research Centre (NIHR203311) and Crohn's & Colitis UK (Leeds, M2023-5 Subramanian). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

Z. Xu and J. Rittscher are with the Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, OX3 7DQ, Oxford, United Kingdom (emails: {ziang.xu, jens.rittscher}@eng.ox.ac.uk)

S. Ali is with the School of Computing, Faculty of Engineering and Physical Sciences, University of Leeds, LS2 9JT, Leeds, United Kingdom (corresponding email: s.s.ali@leeds.ac.uk)

I. Introduction

IMAGE classification, detection, and segmentation tasks have been extensively studied by the biomedical image analysis community [1]. 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) [2], X-ray [3], magnetic resonance imaging (MRI) [4] and endoscopy [5].

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. [6]. 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 [7]. Data curation challenges are thus harder to tackle, leading only to sub-optimal results in supervised learning frameworks [8]. Several studies have also found that most supervised methods lead to a huge performance drop when applied to different centre datasets [8]. 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.

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 [9]. In the medical imaging field, SSL has been used extensively for different tasks, including disease classification [3], [10], lesion

region detection [4], [11] and segmentation [12], [13]. 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, [4] 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 [5] 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 focus on two different lesions found in the colon and rectum, and we aim to devise a robust SSL-based approach to build automated techniques with CNN-based networks. To this end, we propose to develop an SSL approach for ulcerative colitis (UC), a chronic intestinal inflammatory disease, and polyps that are precursor lesions for colorectal cancer. UC is a severe medical condition and requires patient monitoring and risk stratification. Patients with UC have an increased risk of developing colorectal cancer and are therefore put under regular colonoscopy surveillance. Gastroenterologists use the Mayo Endoscopic Score [14] (MES, see Fig. 1 (on the left)), a widely accepted predictive indicator for malignant transformation in UC, as a classification task based on visual appearances. Similarly, polyps that are addressed in this work as detection and segmentation downstream tasks are precursor lesions [15]. Large and cancerous polyps are resected during clinical surveillance itself. However, optimally localising and segmenting the polyps can help these clinical procedures. Automated classification, detection, and segmentation methods can help reduce missed operator variability in these procedures and be helpful in transforming patient care and management.

Supervised learning methods struggle to learn a feature representation that discriminates between the different categories even if trained on large, labelled datasets. The data presented in Fig. 1 illustrates this problem in the context of ulcerative colitis scoring. After supervised learning, we can still observe significant confusion between the different classes. In this work, we propose a novel self-supervised learning strategy for endoscopic image analysis, referred to as "SSL-CPCD". Our approach is based on novel ideas on combining loss functions both at the single instance-level and group-level instance (i.e., clustered samples with similar representations are classified using the k-means clustering approach into a specific class or instance) using image frames and patch-level representations. The proposed losses are used in a pretext-invariant representation learning (PIRL) [16] context but here we utilise patch-level and image-level representations that are learnt at single and grouped instances (i.e., clustered using k-means), amplifying the power of learning discriminative features. For loss functions, unlike classical Noise Contrastive Estimation (NCE) [17], we exploit the additive angular margin loss [18] technique that has proven to pull apart negative samples from positives. In this work, we introduce the additive angular margin loss within the NCE loss at both patch- and image-level instances. Finally, we perform instance-group discrimination from k-means clustered instances similar to [19], but these are performed at both patch-level and image-level. Jointly, we refer to this loss as a composite pretext-class discrimination loss (CPCD). It is to be noted that SSL techniques only require image transformations during pretext training but do not require complex transformations during the fine-tuning stage for downstream tasks. Thus, this two-stage training process enables models to learn robust representations that are further disentangled and refined towards the downstream task with limited data, making it very suitable for endoscopic image analysis.

Compared to pretext-invariant representations (PIRL) approach [16] we introduce a new contrastive estimation block (CEM-block) and an unsupervised $k-$means clustering block [19] for discrimination between both patches and image-level instances which is also different to our previous work where we used PIRL with patch-level discrimination only (referred to as PIRL-PLD) [20]. CEM block introduces computation of contrastive loss with an added angular margin [18] to increase the separation between target embedding and negative samples, providing stronger discriminative power to the NCE loss. Unlike our previous work [21] which used arc-cosine loss with additive angular margin [18] in the downstream task, we have integrated the additive angular margin loss in the pretext task learning stage itself to exploit better representations. In addition, our current setup does not require network and loss changes in the fine-tuning stage. Similarly, for our clustering block unlike Wang et al. [19], our approach consists of grouping-across-view approach with both instance-level and group-level discrimination in not only single images but also with patches. Salient features in endoscopic images that identify a specific disease or lesion is primarily localised and usually not distinctive from surrounding areas. Learning global and local features and their association can be crucial in distinguishing them from the surrounding mucosa, i.e., understanding global and local changes in patterns associated with a disease or lesion. For instance, in colonoscopy, the phenotypic appearances of ulcerative colitis (as illustrated in Fig. 1) highlight the need for both image-level (i.e., global) and patch-level (i.e., local) discrimination. In addition, due to the mucosal property observed in endoscopy, more comprehensive loss functions that incorporate both local and global feature separations are required, which makes our approach unique, more accurate, and robust to different lesion types.

In this work, we propose a novel approach with a comprehensive loss function that improves learning on the pretext task by using both image-level and patch-level discrimination. To this extent, we also use memory banks [22] 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 [20], we only explored classification task for downstream task with single disease type and used NCE loss similar to PIRL together with added patch level dis-

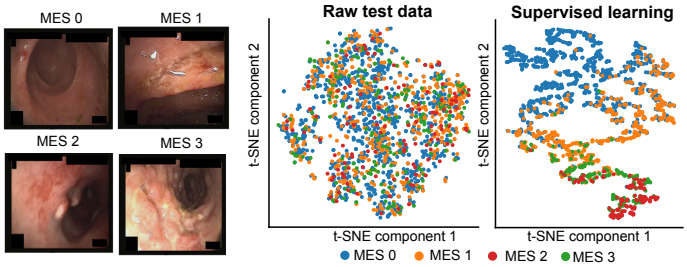

Fig. 1. Endoscopic image analysis for ulcerative colitis scoring. (on left) Representative images for Mayo endoscopic scoring (MES) from 0 to 3, and (on right) t-SNE plots [27] with perplexity of 15 and maximum iterations of 1000 for all test samples before learning and after supervised learning using ResNet50 [28].

crimination 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 [23] in our network for further improvement. All data descriptions and code are available at `https://github.com/EricXuziang/SSL-CPCD`. 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 [18] 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 [24], CVC-ClinicDB [25], LIMUC [26], and our in-house dataset.
- We show that our SSL-CPCD-based method outperforms several SOTA SSL strategies by a large margin.

II. Related work

A. Deep learning in gastrointestinal endoscopy

1) Classification task: Ulcerative colitis (UC) scoring is based on Mayo Endoscopic Scoring (MES) in clinical decision-making. Several CNN-based architectures have been proposed to automate MES. For example, Stidham et al. [29] used an Inception V3 model to train and evaluate MES scores in still endoscopic frames where they used 16k UC images and obtained an accuracy of 67.6%, 64.3% and 67.9% for the three MES classes. Recently, Mokter et al. [30] proposed a method to classify UC severity in colonoscopy videos by detecting vascular (vein) patterns using three CNN networks and a training dataset comprising over 67k frames. Similarly, Ozawa et al. [31] used a CNN for binary classification only to elevate the problem of poor accuracies across classes and used still frames comprising 26k training images, which first between normal (comprising of MES 0 and MES 1) while next class as combined moderate (MES 2) and severe (MES 3). Gutierrez et al. [32] also used the CNN model to predict only a binary version of the MES scoring.

2) Detection task: Polyp detection task has been more widely researched compared to UC classification. Lee et al. [33] used YOLOv2 and validated the algorithm on public datasets and colonoscopy videos, demonstrating real-time capability as one of the milestones. Zhang et al. [34] 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. [35], and Shin et al. [36] used Mask R-CNN and Faster RCNN with different backbones to detect polyps, respectively. While these methods achieve high precision, they fall short in delivering real-time performance.

3) Segmentation task: Polyp segmentation task is the most widely researched topic in endoscopic image analysis. Zhou et al. [37] proposed a technique called U-Net++ based on U-Net, which fully utilises multi-scale features to obtain superior results. Fan et al. [38] proposed a parallel inverse attention-based network (PraNet). PraNet employs a partial decoder to aggregate features in high-level layers and mine boundary cues using an inverse attention module. A Shallow Attention Network (SANet) was proposed by [39]. SANet used a colour swap operation to decouple image content and colour and force the model to pay more attention to the shape and structure of the object. Recently, Srivastava et al. [40] proposed a Multi-Scale Residual Fusion Network (MSRF-Net). MSRF-Net can exchange multi-scale features of different receptive fields using dual-scale dense fusion (DSDF) blocks.

B. Attention mechanism

Attention can make the model more focused, extract the most relevant features, and ignore irrelevant information. It also overcomes the size limitation of the receptive field and can focus on the contribution of global features to the current region [41]. Attention-based models have achieved state-of-the-art performance in medical images such as skin cancer, endoscopy, CT, and X-ray (Sinha and Dolz [42], Zhao et al. [43], Kaul et al. [2], Gu et al. [44]). Zhao et al. [43] proposed an adaptive cosine similarity network with a self-attention module to automatically classify gastrointestinal endoscope images. The self-attention block replaces the conv+BN/Relu operation in traditional CNN and uses a cosine-based self-adapting loss function to adjust the scale parameters automatically, achieving 95.7% on average accuracy in the wireless capsule endoscopy dataset.

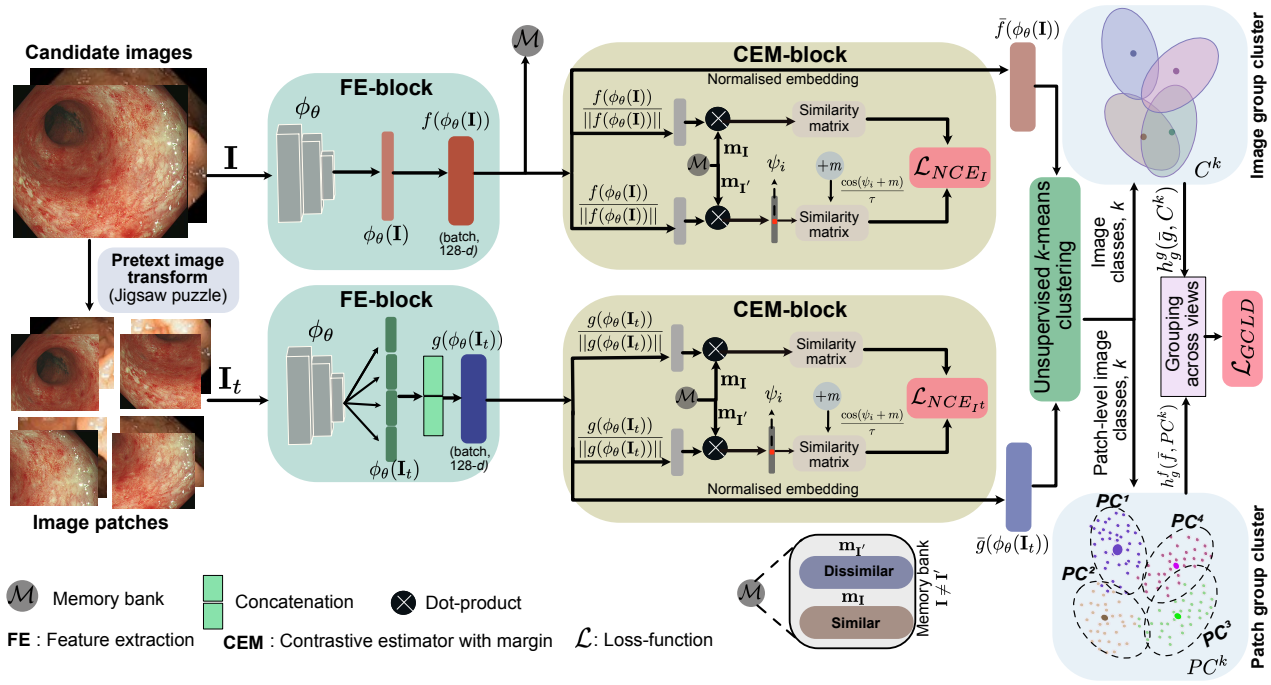

Fig. 2. **Block diagram of our proposed self-supervised learning framework with composite pretext-class discrimination losses (SSL-CPCD).** ResNet50 encoder network [28] is fine-tuned with original images with transformations and randomly shuffled patches as patch transformation mimicking shuffled jigsaw puzzle as in [16] in a self-supervised setting to enable semantically meaningful representation learning for improved generalisability and accuracy in downstream tasks. Contrastive estimator with margin (CEM-block) is separately computed for image-level and patch-level instances. Further, a group-wise contrastive loss is computed by comparing the centroids at patch-level (PC^k) group and at image-level (C^k). Memory bank $\mathcal{M}$ [22] is used for storing all representations.

C. 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. 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. In SSL, the pretext task typically applies a transformation to the input image and predicts the properties of the transformation from the transformed image. During pretext learning, a single model [16], [45], [46] 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. Chen *et al.* [45] proposed the SimCLR model, which 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. [46] 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 methods that encourage the construction of covariant image representations to the transformations, pretext-invariant representation learning (PIRL) [16] 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 [47] was used as a image patch representation for PIRL representation learning. Wang et al. [19] 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.

In recent years, self-supervised learning has also been applied in the field of medical image analysis but not much on the endoscopic image analysis. Azizi et al. [3] used multi-instance contrastive learning based on self-supervision on medical images, followed by a fully supervised fine-tuning method for the final classification of available task-specific losses. They improved top-1 accuracy by 6.7% and 1.1% on dermatology and chest X-ray classification, respectively. Zeng et al. [12] proposed SeSe-Net for medical image segmentation. SeSe-Net is divided into two neural networks, ”worker” and ”supervisor”. In the first stage, the standard data set is learned and segmented, and a training set is generated, and then the supervisor further supervises the learning process in the second stage so that the worker further improves the performance on the non-labelled dataset. Chen et al. [4] proposed a novel self-supervised learning strategy based on context restoration to

change the spatial information of an image by selecting and exchanging two patches in the same image to learn enough pronounced semantic representations. It was validated on 2D fetal ultrasound images, abdominal computed tomography images, and brain magnetic resonance images. Recently, Ciga et al. [13] used a residual network pre-trained with self-supervised learning to learn generalisable features and then used the pre-trained network in downstream tasks to perform multiple tasks on multiple multi-organ digital histopathology datasets. Similarly, SSL methods (including SimCLR and MoCo) was recently used for the phase recognition and tool presence detection tasks in the surgical domain [48].

III. 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. 2) - firstly, exploiting positive and negative samples for noise contrastive loss estimation ($\mathcal{L}_{\text{NCE}}$). Unlike classically used $\mathcal{L}_{\text{NCE}}$ [17], we integrate an added angular margin [18] 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 [19] 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., $< \bar{g}_i, C_i^k >$, and between patch-level cluster centroid PC_i^k and normalised image-level feature embedding $\bar{f}_i$, i.e., $(< \bar{f}_i, PC_i^k >)$. This determines how close the embedding of the same class k (k-means clustered) are is in either image instances C_i^k or patch instances PC_i^k. Thus, 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 $\mathcal{L}_{\text{PC}}^k$ and image level $\mathcal{L}_{\text{C}}^k$ 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 [22] 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. 2.

A. Feature extraction (FE) block

Let the endoscopy dataset $\mathcal{D}$ consist of N image samples, denoted as $\mathcal{D} = \{\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_N\}$. We use a set of image transformations $\mathcal{T}$ to create and reshuffle m number of image patches for each image in the dataset ($\mathcal{D}$, $\mathcal{P} = \{\mathbf{I}_{1t}^1, ..., \mathbf{I}_{1t}^m, ..., \mathbf{I}_{Nt}^1, ..., \mathbf{I}_{Nt}^m\}$ 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 [28]) with free parameters θ 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 $f(.)$ 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 [16]. But unlike solving a jigsaw pretext task [47], 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 64×64 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 $g(.)$ [45] is applied to re-scale the representations to a 128-dimensional feature vector.

Memory banks: The memory bank [22] $\mathcal{M}$ 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 $\mathbf{m_I}$ with assigned indexes that helps to build negative samples $\mathbf{m_{I'}}$ 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 [16].

B. Contrastive loss estimation with margin (CEM) block

Noise contrastive estimator (NCE) [16], [17], [49] 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 [16], [17], the NCE loss models the posterior probability $h(f_i, f_i^+)$ of sample i given a data distribution $\mathcal{D}$ and temperature parameter τ as:

$$h(f_i, f_i^+) = \left(\frac{\exp \frac{<f_i, f_i^+>}{\tau}}{\exp \frac{<f_i, f_i^+>}{\tau} + \sum_{f_i^- \in \mathcal{D}_n} \exp \frac{<f_i, f_i^->}{\tau}} \right) \quad (1)$$

However, unlike [16], [17], we propose to add an angular margin [18] 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 (ψ) $\psi = \arccos < f_i, \mathbf{m}_{\mathbf{I}'} >$ and then add an angular margin m to the computed angle, i.e. $\psi_{new} = \psi + m$. Finally, *cosine* of the ψ_{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. 2 represents the posterior probability with added angular margin [18]. Here, the angular additive margin aims to further increase the separation between feature representations of target embedding and negative samples.

$$h^{new}(f_i, f_i^+) = \left(\frac{\exp \frac{<f_i, f_i^+>}{\tau}}{\exp \frac{<f_i, f_i^+>}{\tau} + \sum_{f_i^- \in \mathcal{D}_n} \exp \frac{\cos(\psi_{new})}{\tau}} \right) \quad (2)$$
$$\text{with } \psi_{new} = \arccos(< f_i, f_i^- >) + m$$

The total NCE loss entails minimising the joint-loss (logarithmic) function of Eq. 2 both at the image-level and patch-level configurations (also see Fig. 2) 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:

$$\mathcal{L}_{NCE}^{total}(\mathbf{I}, \mathbf{I}_t) = \lambda \mathcal{L}_{NCE_I}(\mathbf{m}_\mathbf{I}, \bar{f}(\phi_\theta(\mathbf{I}))) + (1-\lambda)\mathcal{L}_{NCE_{I_t}}(\mathbf{m}_\mathbf{I}, \bar{g}(\phi_\theta(\mathbf{I}_t))). \quad (3)$$

Each NCE loss component can be established as a negative log-posterior distribution of data samples and negative samples [16], [17], [22]. 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. 2) using posterior probability in Eq. 2 given as below [22]:

$$\mathcal{L}_{NCE_I}(\mathbf{m}_\mathbf{I}, \bar{f}(\phi_\theta(\mathbf{I}))) = -\log[h^{new}(\bar{f}(\phi_\theta(\mathbf{I})), \mathbf{m}_\mathbf{I})] - \sum_{\mathbf{I}' \in \mathcal{D}_n} \log[1 - h^{new}(\mathbf{m}_{\mathbf{I}'}, \bar{f}(\phi_\theta(\mathbf{I})))], \quad (4)$$

$$\& \quad \mathcal{L}_{NCE_{I_t}}(\mathbf{m}_\mathbf{I}, \bar{g}(\phi_\theta(\mathbf{I}_t))) = -\log[h^{new}(\bar{g}(\phi_\theta(\mathbf{I}_t)), \mathbf{m}_\mathbf{I})] - \sum_{\mathbf{I}' \in \mathcal{D}_n} \log[1 - h^{new}(\mathbf{m}_{\mathbf{I}'}, \bar{g}(\phi_\theta(\mathbf{I}_t)))]. \quad (5)$$

Since, we take the logarithmic of function in Eq. 2, 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 $\mathbf{m}_\mathbf{I}$ that will damp the parameter updates in the weights ϕ_θ. It also further penalises the representations at both image and patch-level from other set of negative images $\mathbf{I}^{'}$.

C. 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 [45] 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 [19]. 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 $k = \{1, ..., n\}$, 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.

D. 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 $\bar{g}(\phi_\theta(\mathbf{I}))$ 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 [17] across the views can be defined using the expression in Eq. (1). 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 [19]:

$$h_g^f(\bar{f}_i, PC_i) = -\log \frac{\exp \frac{<\bar{f}_i, PC_i>}{\tau}}{\exp \frac{<\bar{f}_i, PC_i>}{\tau} + \sum_{j \neq i} \exp \frac{<\bar{f}_i, PC_j>}{\tau}} \quad (6)$$

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

$$h_g^g(\bar{g}_i, C_i) = -\log \frac{\exp \frac{<\bar{g}_i, C_i>}{\tau}}{\exp \frac{<\bar{g}_i, C_i>}{\tau} + \sum_{j \neq i} \exp \frac{<\bar{g}_i, C_j>}{\tau}} \quad (7)$$

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

$$\mathcal{L}_{GCLD}=\sum_{k=1}^{k}\sum_{i=1}^{N}\{\lambda h_g^f(\bar{f}_i, PC_i^k)+(1-\lambda)h_g^g(\bar{g}_i, C_i^k)\} \quad (8)$$

Each patch-level and image-level instance are already compared independently within our CEM-block (see Fig. 2 and Eq. 3) 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. 8 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.

E. Proposed CPCD 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 estimation loss refers to the single instance-level representations and the group-wise cross-level discrimination loss. Empirically, λ in Eq. 3 and Eq. 8 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. 3 and Eq. 8 with $\lambda^{'}$ as weighting factor and is given as:

$$\mathcal{L}_{CPCD}=\sum_{k=1}^{k}\lambda^{'}\underbrace{\sum_{i=1}^{N}0.5\cdot\{h_g^f(\bar{f}_i, PC_i^k)+h_g^g(\bar{g}_i, C_i^k)\}}_{\mathcal{L}_{GCLD}}+$$
$$(1-\lambda^{'})\underbrace{\sum_{i=1}^{N}0.5\cdot\{\mathcal{L}_{NCE_I}(\mathbf{m_I}, \bar{f}_i(\phi_\theta(\mathbf{I})))+\mathcal{L}_{NCE_{I^t}}(\mathbf{m_I}, \bar{g}_i(\phi_\theta(\mathbf{I}^t))\}}_{\mathcal{L}_{NCE}} \quad (9)$$

IV. Experiments

A. Dataset and setup

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 [26] and one in-house dataset (collected under universal patient consenting at the Translational Gastroenterology Unit, John Radcliffe Hospital, Oxford) are applied. Kvasir-SEG [24] and CVC-ClinicDB [25] are used for the segmentation task. Similarly, Kvasir-SEG [24] for experiments on polyp detection as a downstream task. In the pretext task for the detection and segmentation tasks, we have used polyp samples from Kvasir-SEG [24] and 1000 non-polyp samples from the SUN dataset [50] 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. 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 I. All datasets are publicly available including the in-house dataset used for generalisability assessment. The in-house dataset can be downloaded at `https://doi.org/10.7303/syn52674005`.

TABLE I
COLONOSCOPIC DATASETS USED IN OUR EXPERIMENTS

Dataset	Images	Input size	Train	Valid	Test
Ulcerative colitis classification					
LIMUC [26]	11276	224×224	8631	959	1686
In-house	251	224×224	0	0	251
Polyp segmentation					
Kvasir-SEG [24]	1000	Variable	800	100	100
SUN(non-polyp) [50]	1000	Variable	1000	0	0
CVC-ClinicDB [25]	612	384×288	0	0	612
Polyp detection					
Kvasir-SEG [24]	1000	Variable	800	100	100
SUN [50]	1000	Variable	1000	0	0

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. 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.

3) Implementation details: The proposed method is implemented using PyTorch [52]. 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 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 224×224 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

TABLE II
QUANTITATIVE COMPARISON FOR UC CLASSIFICATION TASK ON LIMUC DATASET

Method	Backbone	Top 1	Top 2	F1	Spec.	Recall	QWK	P-values
Baseline [28]	R50	0.7532	0.9346	0.6689	0.8630	0.7011	0.8237	7.022e-07
Baseline [23], [28]	R50-Att.	0.7556	0.9414	0.6727	0.8692	0.7029	0.8290	4.342e-06
SimCLR [45]	R50	0.7355	0.9387	0.6631	0.8510	0.6752	0.8083	6.527e-11
SimCLR [45]	R50-Att.	0.7384	0.9219	0.6649	0.8431	0.6942	0.8102	4.843e-09
SimCLR+DCL [51]	R50	0.7555	0.9450	0.6729	0.8635	0.6897	0.8269	1.506e-07
SimCLR+DCL [51]	R50-Att.	0.7568	0.9367	0.6755	0.8669	0.6952	0.8287	8.760e-07
MoCoV2+CLD [19]	R50	0.7574	0.9493	0.6788	0.8721	0.6959	0.8309	7.691e-12
MoCoV2+CLD [19]	R50-Att.	0.7598	0.9536	0.6812	0.8709	0.7047	0.8333	6.787e-09
PIRL [16]	R50	0.7651	0.9637	0.6859	0.8874	0.7098	0.8376	4.201e-04
PIRL [16]	R50-Att.	0.7740	0.9610	0.6918	0.8893	0.7133	0.8460	1.621e-03
PIRL+PLD (ours) [20]	R50	0.7752	0.9666	0.7040	0.8891	0.7129	0.8509	6.043e-03
PIRL+PLD (ours) [20]	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	**0.9043**	0.7198	0.8693	-
SSL-CPCD (ours)	R50-Att.	**0.7977**	**0.9750**	**0.7279**	0.9008	**0.7259**	**0.8746**	-

a batch size of 32, and a stopping criteria with the patience of 20 epochs. 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. 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°), random horizontal flip (probability, $p = 0.5$), random crop of input size with padding adding of 10 pixels, random affine transformation (10°) 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 (please refer to the fine-tuning codes available at `https://github.com/EricXuziang/SSL-CPCD`).

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% 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. (8)), 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 [16] has been used with the same hyperparameters. For Eq. (3) we use $\lambda = 0.5$ and use $\tau = 0.4$ for computing the function $h(.,.)$ in Eq. (1, 2, 6, and 7). 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 IV-D.2. Hyperparameter tuning for our SSL parameters (τ, and λ) 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.

B. Results

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

1) Comparison for UC classification task: ResNet50 [28] (R50) and ResNet50 with convolution-block attention module [23] (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 II 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 [45], SimCLR+DCL [51], MoCoV2+CLD [46] and PIRL [16] 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.

2) 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 [53] as the baseline network for both the supervised and the self-supervised learning approaches. The quantitative results from Table III 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,

TABLE III
QUANTITATIVE COMPARISON FOR POLYP DETECTION TASK USING KVASIR-SEG DATASET

Method	Backbone	mAP	AP25	AP50	AP75	APsmall	APmedium	APlarge	P-values
RetinaNet [53]	R50	0.8637±0.101	0.9377	0.8965	0.6973	0.4832	0.7507	0.8398	6.150e-03
RetinaNet [53]	R50-Att.	0.8729±0.082	0.9436	0.9097	0.7045	0.4871	0.7603	0.8419	1.263e-03
SimCLR [45]	R50	0.8501±0.090	0.9259	0.8837	0.6709	0.4641	0.7416	0.8237	1.337e-05
SimCLR [45]	R50-Att.	0.8532±0.076	0.9278	0.8818	0.6846	0.4679	0.7403	0.8302	9.270e-03
SimCLR+DCL [51]	R50	0.8537±0.087	0.9269	0.8853	0.6852	0.4709	0.7429	0.8321	5.057e-04
SimCLR+DCL [51]	R50-Att.	0.8592±0.101	0.929	0.8893	0.6887	0.4739	0.7467	0.8403	2.156e-04
MoCoV2+CLD [19]	R50	0.8457±0.072	0.9273	0.9028	0.6845	0.4779	0.7491	0.8346	1.869e-09
MoCoV2+CLD [19]	R50-Att.	0.8519±0.070	0.9412	0.9044	0.7019	0.4859	0.7563	0.8402	7.792e-07
PIRL [16]	R50	0.8612±0.078	0.9403	0.8931	0.6929	0.4839	0.7487	0.8317	2.707e-02
PIRL [16]	R50-Att.	0.8677±0.073	0.9408	0.8961	0.702	0.4863	0.7589	0.8408	8.771e-03
PIRL+PLD (ours) [20]	R50	0.8649±0.077	0.9431	0.9027	0.7042	0.4919	0.7547	0.8426	5.694e-03
PIRL+PLD (ours) [20]	R50-Att.	0.8738±0.071	0.9466	0.9053	0.7069	0.4977	0.7601	0.8458	1.178e-02
SSL-CPCD (ours)	R50	0.8709±0.070	0.9421	0.9192	0.7107	0.5033	0.763	0.8542	-
SSL-CPCD (ours)	R50-Att.	**0.8862±0.067**	**0.9469**	**0.9227**	**0.7197**	**0.5105**	**0.7703**	**0.8598**	-

TABLE IV
QUANTITATIVE COMPARISON FOR POLYP SEGMENTATION TASK

Method	Backbone	DSC	F2	Recall	PPV	P-values
U-Net [54]	none	0.7933±0.053	0.7671	0.7945	0.9131	2.881e-10
Res-UNet [55]	R50	0.7867±0.074	0.7667	0.7723	0.9139	1.670e-04
Res-UNet [55]	R50-Att.	0.792±0.066	0.7743	0.7862	0.9187	7.117e-06
SimCLR [45]	R50	0.7892±0.068	0.7621	0.7639	0.9146	2.936e-07
SimCLR [45]	R50-Att.	0.7945±0.058	0.7759	0.7903	0.9162	6.977e-05
SimCLR+DCL [51]	R50	0.7879±0.056	0.7609	0.7653	0.9169	4.690e-05
SimCLR+DCL [51]	R50-Att.	0.7933±0.051	0.7822	0.7741	0.9038	8.816e-06
MoCoV2+CLD [19]	R50	0.7946±0.093	0.779	0.7846	0.9173	3.697e-11
MoCoV2+CLD [19]	R50-Att.	0.8029±0.077	0.7953	0.7998	0.9201	7.327e-09
PIRL [16]	R50	0.7906±0.069	0.7842	0.7946	0.9135	7.502e-03
PIRL [16]	R50-Att.	0.7969±0.067	0.7893	0.8056	0.9177	1.432e-02
PIRL+PLD (ours) [20]	R50	0.8069±0.058	0.7913	0.8093	0.9189	5.519e-03
PIRL+PLD (ours) [20]	R50-Att.	0.8116±0.049	0.7989	0.8132	0.9211	4.107e-02
SSL-CPCD (ours)	R50	0.8173±0.047	0.8032	0.8104	0.9217	-
SSL-CPCD (ours)	R50-Att.	**0.8232±0.043**	**0.8081**	**0.8234**	**0.9259**	-

TABLE V
GENERALISATION STUDY FOR THE UC CLASSIFICATION TASK

Method	Backbone	Top 1	Spec.	Recall	QWK	P-values
Baseline [28]	R50	0.5856	0.7239	0.5569	0.5379	2.315e-11
Baseline [23], [28]	R50-Att.	0.6055	0.7539	0.5739	0.6572	6.812e-08
SimCLR [45]	R50	0.5737	0.7020	0.5256	0.5611	6.076e-09
SimCLR [45]	R50-Att.	0.5777	0.7139	0.5420	0.6018	1.497e-06
SimCLR+DCL [51]	R50	0.5976	0.7297	0.5622	0.6345	4.277e-07
SimCLR+DCL [51]	R50-Att.	0.6016	0.7458	0.5758	0.6542	9.603e-06
MoCoV2+CLD [19]	R50	0.6175	0.7716	0.5878	0.6939	5.657e-08
MoCoV2+CLD [19]	R50-Att.	0.6135	0.7823	0.5737	0.6902	6.069e-11
PIRL [16]	R50	0.6255	0.8213	0.6097	0.7312	1.370e-03
PIRL [16]	R50-Att.	0.6335	0.8397	0.6139	0.7469	9.677e-04
PIRL+PLD (ours) [20]	R50	0.6370	0.8359	0.6197	0.7507	2.574e-02
PIRL+PLD (ours) [20]	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.	**0.6733**	**0.8677**	**0.6403**	**0.7887**	-

our method is better on mean average mAP but significantly improves over AP50, AP75 and size-based metrics.

3) 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 [55] as the baseline network for both the supervised and the self-supervised learning approaches. Results for supervised U-Net [54] approach have also been provided for comparison. Table IV 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.

4) 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 paper 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).

C. 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 different institutions.

1) Generalisibility 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 V 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 V shows that our SSL-CPCD outperforms other SOTA methods in various evaluation metrics.

2) 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 VI. Our proposed SSL-CPCD drastically surpassed baseline supervised approaches (over 10% on DSC for U-Net and over 7% on DSC

TABLE VI
GENERALISATION STUDY FOR SEGMENTATION TASK

Method	Backbone	DSC	F2-score	Recall	PPV	P-values
U-Net [54]	none	0.5826±0.057	0.6029	0.5942	0.7633	1.009-e12
Res-UNet [55]	R50	0.6092±0.040	0.6379	0.6265	0.8218	8.708-e05
Res-UNet [55]	R50-Att.	0.6027±0.072	0.6499	0.6372	0.8065	6.363-e04
SimCLR [45]	R50	0.5942±0.069	0.6498	0.6334	0.8312	1.982-e07
SimCLR [45]	R50-Att.	0.6113±0.055	0.6556	0.6673	0.8329	6.621-e04
SimCLR+DCL [51]	R50	0.6039±0.059	0.6501	0.6586	0.8293	8.193-e06
SimCLR+DCL [51]	R50-Att.	0.6092±0.049	0.6679	0.6598	0.8301	2.770-e03
MOCOv2+CLD [19]	R50	0.6268±0.047	0.6498	0.6509	0.8277	9.913-e09
MOCOv2+CLD [19]	R50-Att.	0.632±0.050	0.6691	0.6675	0.8362	7.076-e07
PIRL [16]	R50	0.6196±0.049	0.6742	0.6703	0.8378	5.501-e03
PIRL [16]	R50-Att.	0.6277±0.053	0.6801	0.6770	0.8396	6.950-e04
PIRL+PLD (ours) [20]	R50	0.6459±0.062	0.6617	0.6802	0.8311	1.669e-03
PIRL+PLD (ours) [20]	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.	**0.6793**±0.040	**0.6978**	**0.6897**	**0.8488**	-

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.).

D. 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.

1) Loss functions: Table VII shows the quantitative results of our ablation study in loss functions. Initially, our proposed method, which contains three loss functions, achieves 79.12% on top 1 accuracy and 72.09% on the F1 score for the classification task. 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. 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, mAP, and DSC improved respectively by 2.61%, 0.97% and 2.07%).

2) Impact of hyper-parameters: The quantitative results for the ablation study of different parameter settings are shown in Table VIII. We set different weight and temperature in Eq. (3-8). 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 VIII, when weight and temperature parameters are 0.5 and 0.4, respectively, our method achieves the best results in classification and segmentation

TABLE VII
ABLATION STUDY RESULTS FOR DIFFERENT LOSS FUNCTIONS ON VALIDATION SET

Loss function	Class. task		Det. task		Seg. task	
	Top 1	F1	AP50	mAP	DSC	PPV
NCE	0.8341	0.7693	0.9527	0.9004	0.8883	0.9438
NCE+GCLD	0.8495	0.781	0.9613	0.9062	0.8987	0.9569
NCE+GCLD+m	**0.8603**	**0.7933**	**0.9697**	**0.9119**	**0.9037**	**0.9604**

TABLE VIII
EFFECT OF DIFFERENT HYPER-PARAMETER SETTING USING VALIDATION SET

Parameter settings		Top 1	AP50	DSC
$\lambda^{'}$	τ			
0.1	0.2	0.8342	0.9461	0.8897
0.25	0.2	0.8415	0.9489	0.8917
0.5	0.2	0.8509	0.9611	0.8968
1	0.2	0.8405	0.9522	0.8903
0.1	0.4	0.8425	0.9503	0.8940
0.25	0.4	0.8467	0.9517	0.8987
0.5	0.4	**0.8603**	0.9653	**0.9037**
1	0.4	0.8446	0.9549	0.8962
0.1	0.6	0.8332	0.9566	0.8863
0.25	0.6	0.8383	0.9607	0.8907
0.5	0.6	0.8498	**0.9697**	0.8935
1	0.6	0.8352	0.9625	0.8886

TABLE IX
HYPER-PARAMETER STUDY FOR DIFFERENT NUMBER OF CLUSTERS

No. of clusters k	Top 1	AP50	DSC
2	0.8533	**0.9697**	**0.9037**
3	0.8540	0.9578	0.9015
4	**0.8603**	0.9618	0.8999
5	0.8582	0.9601	0.8957
7	0.8498	0.9627	0.9004
10	0.8519	0.9609	0.8986

TABLE X
ABALATION STUDY FOR DIFFERENT LEARNING RATE USING VALIDATION SET

Method	LR	Top 1	F1	Spec.	Recall	QWK
SimCLR [45]	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	**0.7433**	0.8751
	0.001	**0.8049**	**0.7345**	**0.9239**	0.7414	**0.8769**
	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 (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
	0.001	**0.8603**	**0.7933**	**0.9523**	**0.7911**	**0.9326**
	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 XI
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 (ResNet50) [28]	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 [16]	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 [20]	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 (ours)	100	**0.7912**	**0.7209**	**0.9043**	**0.7198**	**0.8693**
	50	**0.7811**	**0.7133**	**0.8977**	**0.7158**	**0.8562**
	20	**0.7692**	**0.6978**	**0.8749**	**0.6897**	**0.8242**
	10	**0.7551**	**0.6897**	**0.8666**	**0.6709**	**0.8001**

TABLE XII
COMPARISION 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 [53]	100	0.8637±0.101	0.9377	0.8965	0.6973	6.150e-03
	50	0.8321±0.123	0.9051	0.8709	0.6779	1.479e-05
	20	0.8207±0.095	0.8778	0.8587	0.6583	3.290e-04
	10	0.8093±0.117	0.8599	0.8327	0.6467	3.157e-06
PIRL [16]	100	0.8612±0.078	0.9403	0.8931	0.6929	2.707e-02
	50	0.8455±0.081	0.9194	0.8793	0.6799	9.670e-05
	20	0.8260±0.087	0.8817	0.8571	0.6642	4.551e-04
	10	0.8135±0.082	0.8627	0.8369	0.6498	7.619e-05
PIRL+PLD [20]	100	0.8649±0.077	0.9431	0.9027	0.7042	5.694e-03
	50	0.8398±0.069	0.9156	0.8769	0.6826	1.191e-03
	20	0.8261±0.072	0.8942	0.8551	0.6717	2.233e-02
	10	0.8172±0.079	0.8653	0.8390	0.6521	8.582e-03
SSL-CPCD (ours)	100	**0.8709±0.070**	**0.9421**	**0.9192**	**0.7107**	-
	50	**0.8587±0.059**	**0.9229**	**0.8917**	**0.7027**	-
	20	**0.8439±0.075**	**0.9060**	**0.8764**	**0.6793**	-
	10	**0.8307±0.077**	**0.8966**	**0.8671**	**0.6762**	-

TABLE XIII
COMPARISION 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 [55] (Supervised)	100	0.7867±0.074	0.7667	0.7723	0.9139	1.670e-04
	50	0.7423±0.063	0.7341	0.7369	0.7579	2.511e-03
	20	0.7167±0.077	0.7062	0.7098	0.6809	9.621e-06
	10	0.6949±0.080	0.6897	0.6907	0.6706	5.825e-06
Polyp-PVT [56] (Supervised)	100	0.9067±0.042	0.8832	0.8897	0.9663	3.238e-02
	50	0.8621±0.053	0.8487	0.8479	0.9409	1.181e-02
	20	0.8539±0.059	0.8376	0.8301	0.9241	6.489e-03
	10	0.8346±0.057	0.8192	0.8221	0.9103	4.670e-05
PIRL [16] with Res-UNet (SSL)	100	0.7906±0.069	0.7842	0.7946	0.9135	7.502e-03
	50	0.7589±0.072	0.7433	0.7476	0.8772	9.017e-05
	20	0.7418±0.078	0.7306	0.7411	0.8591	1.109e-06
	10	0.7273±0.086	0.7139	0.7186	0.8267	8.940e-08
PIRL [16] with Polyp-PVT (SSL)	100	0.8997±0.051	0.8781	0.8792	0.9580	2.209e-03
	50	0.8660±0.045	0.8465	0.8518	0.9433	4.091e-05
	20	0.8541±0.059	0.8329	0.8396	0.9387	4.672e-04
	10	0.8302±0.060	0.8201	0.8254	0.9312	2.063e-07
PIRL+PLD [20] with Res-UNet (SSL)	100	0.8069±0.058	0.7913	0.8093	0.9189	5.519e-03
	50	0.7749±0.053	0.7623	0.7646	0.8855	3.431e-07
	20	0.7495±0.067	0.7541	0.7372	0.8661	7.073e-06
	10	0.7371±0.069	0.7230	0.7287	0.8415	2.901e-05
PIRL+PLD [20] with Polyp-PVT (SSL)	100	0.9083±0.049	0.8862	0.9007	0.9624	1.720e-03
	50	0.8691±0.052	0.8513	0.8577	0.9475	2.098e-02
	20	0.8571±0.055	0.8381	0.8431	0.9406	7.574e-03
	10	0.8399±0.063	0.8279	0.8316	0.9344	6.605e-05
SSL-CPCD with Res-UNet (ours)	100	0.8173±0.047	0.8032	0.8104	0.9218	-
	50	0.8063±0.044	0.7867	0.7903	0.8867	-
	20	0.7749±0.050	0.7647	0.7727	0.8803	-
	10	0.7657±0.058	0.7562	0.7597	0.8747	-
SSL-CPCD with Polyp-PVT (ours)	100	**0.9131±0.039**	**0.8960**	**0.9041**	**0.9693**	-
	50	**0.8793±0.037**	**0.8702**	**0.8833**	**0.9567**	-
	20	**0.8697±0.040**	**0.8499**	**0.8669**	**0.9508**	-
	10	**0.8542±0.047**	**0.8407**	**0.8486**	**0.9387**	-

TABLE XIV
COMPARISON WITH LARGER BASELINE BACKBONES FOR CLASSIFICATION, DETECTION, AND SEGMENTATION TASKS.

Method	Parameters	Top 1	AP50	DSC
EfficientNet-v2 [57] (Baseline)	24M	0.7562	0.9090	0.7957
EfficientNet-v2 [57] (+ SSL-CPCD)	24M	0.7948	0.9221	0.8277
EfficientNet-B6 [58] (Baseline)	43M	0.7556	0.9028	0.7911
EfficientNet-B6 [58] (+ SSL-CPCD)	43M	0.7953	0.9207	0.8203
ResNet101 [28] (Baseline)	44.5M	0.7550	0.8993	0.7890
ResNet101 [28] (+ SSL-CPCD)	44.5M	0.7930	0.9198	0.8149
ResNet152 [28] (Baseline)	60.2M	0.7568	0.9047	0.7938
ResNet152 [28] (+ SSL-CPCD)	60.2M	0.7972	0.9214	0.8229
SENet-154 [59] (Baseline)	440M	0.7612	0.9138	0.8016
SENet-154 [59] (+ SSL-CPCD)	440M	0.8021	0.9343	0.8353

tasks with 79.77% on Top 1 accuracy and 82.32% on DSC, respectively. For the detection task, the best performance of our SSL-CPCD was obtained when $\lambda^{'} = 0.5$ and $\tau = 0.6$. We additionally explored different learning rates for all compared methods, and our experiments were conducted on a validation set that suggested the same learning rate of 0.001 (Table X). To substantiate the effect of the number of clusters on cross-level grouping for the loss computation, we conducted an additional study (Table IX). It can be observed that the best results were obtained when using the same number of clusters as the number of classes in each downstream task.

3) Limited data settings: It can be observed that compared to the baseline method and most accurate SSL technique (PIRL [16] and PIRL+PLD [20]), our approach outperforms in all data settings for all cases. For example, in classification task (Table XI) 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. Similarly, for detection task (Table XII) under all data settings, our method outperforms all other approaches with 2%-3% improvements. Finally, for the segmentation downstream task (Table XIII) where two fully supervised models are used residual U-Net (Res-UNet [55]) and polyp segmentation with pyramid vision transformers (Polyp-PVT [56]), 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%).

4) Comparison with other larger models: It can be observed from Table XIV that our approach (SSL-CPCD) consistently provides a perfomance boost of nearly 4% in classification, $\approx 2\%$ in detection, and $\approx 3\%$ in segmentation independent of model size.

5) Model efficiency: Compared to the baseline SSL PIRL model [16], 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).

E. Qualitative Analysis

In the UC classification task in Fig. 3, 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. 3 a) that more compact and interpretable clusters were obtained using perplexity of 15 compared to perplexity of 5. It can be observed (Fig. 3 a.ii)

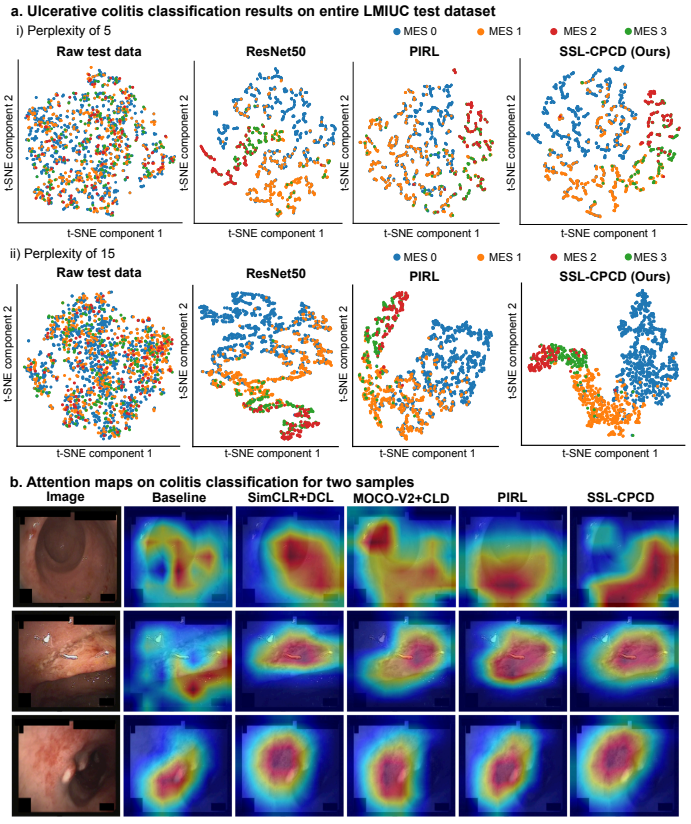

Fig. 3. a) t-SNE plot for the raw test data, baseline network (supervised) and two SSL approaches with perplexity of 5 in a. i) and 15 in a. ii); and b) attention maps of the proposed SSL-CPCD compared to other SOTA methods for multi-class ulcerative colitis classification task (MES 0, MES 1, and MES 2). Principal Components Analysis (PCA) was used to reduce the dimension of the feature maps to 50 as suggested in [27]. 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.

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 [27]. 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 seperation 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.

Similarly, while looking at the attention (Fig. 3 b), the baseline method focuses on the wrong location in some images (see the first and second rows in Fig. 3 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.

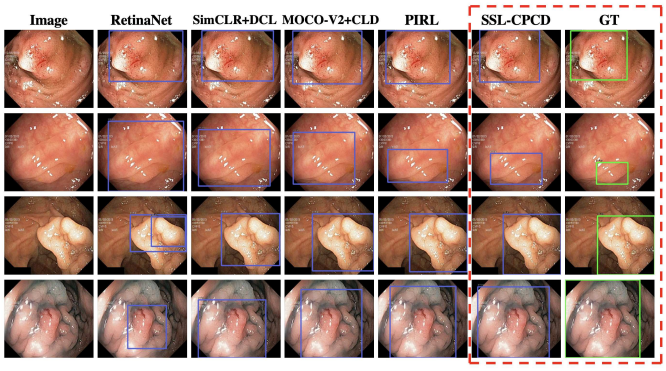

Fig. 4. Qualitative comparison of our proposed SSL-CPCD with other SOTA methods for polyp detection tasks.

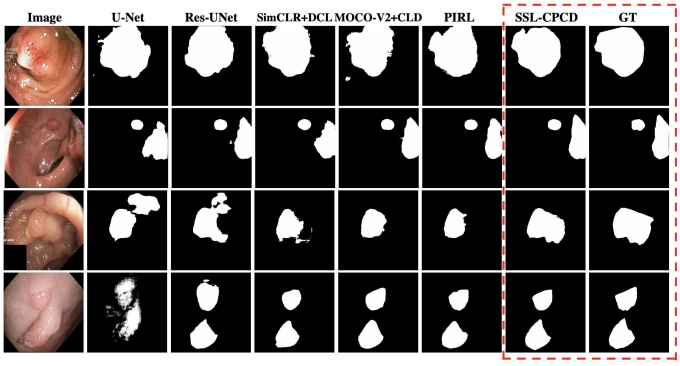

Fig. 5. Qualitative comparison of our proposed SSL-CPCD with other SOTA methods for polyp segmentation task.

For the polyp detection task (Fig. 4), 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), 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.

V. Discussion 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 [60], [61]. 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 [6]. 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 [6]. 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.* [6]. 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 [16], [45], [46], 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.

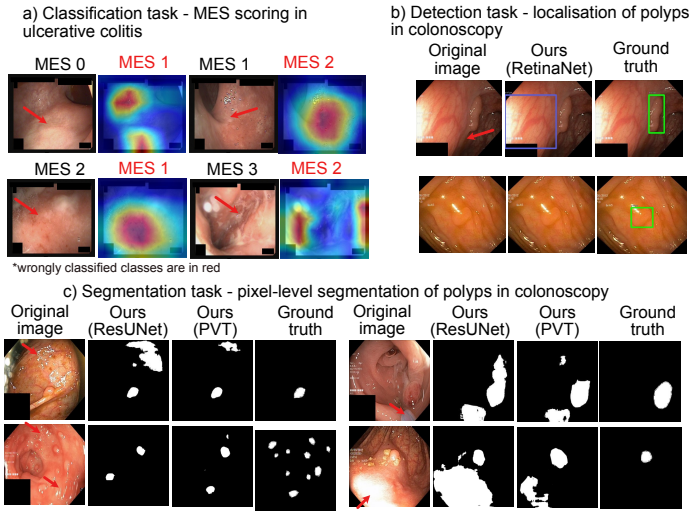

Fig. 6. Limitations of our proposed approach (SSL-CPCD) for different data conditions, including blur, difficult scene views due to organ topology or camera positions, imaging artefacts such as specularity and pixel saturation.

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 VII). Our proposed SSL-CPCD significantly improves all three 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 II). 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 III). Furthermore, we report the best DSC, recall and PPV for the polyp segmentation task on the Kvasir-SEG dataset (see Table Table IV). Furthermore, SSL-CPCD on the generalisability assessment it achieves the highest Top 1 accuracy and QWK of 67.33% and 78.87% (see Table V), and the highest DSC of 67.93% (see Table VI). From Fig. 5, 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. Additionally, P-values from paired t-tests show statistical significance in results from our methods compared to others. It can be observed that for all tasks our approach provided P-value ≤ 0.05 (significantly different, Table II, Table III, Table IV). It can also be observed that the standard deviation provided for main metrics in detection (mAP, Table III) and segmentation (DSC, Table III) is smaller compared to all other methods.

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 6). It can be observed that for the classification task in Figure 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. For the detection task in Figure 6 (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 6 (c). 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 [3]. However, it is essential to note that our SSL-CPCD approach for classification with only 10% data samples (Table XI) can get the performance of fully supervised learning approaches that require 100% labelled data. A similar trend can be observed for detection (Table XII) and segmentation (Table XIII) 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, to understand the effect of cluster size k (Table IX) and the various other hyperparameters (Table VIII), we used a heuristic approach on all SSL approaches for each downstream task, and the best values were identified and reported in the paper. 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.

REFERENCES

[1] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[2] C. 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.

[3] S. 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.

[4] L. Chen *et al.*, "Self-supervised learning for medical image analysis using image context restoration," *Medical image analysis*, vol. 58, p. 101539, 2019.

[5] S. Ali, "Where do we stand in ai for endoscopic image analysis? deciphering gaps and future directions," *npj Digital Medicine*, vol. 5, no. 1, p. 184, Dec 2022.

[6] H. Liu, J. Z. HaoChen, A. Gaidon, and T. Ma, "Self-supervised learning is more robust to dataset imbalance," in *International Conference on Learning Representations*, 2022.

[7] M. Lux and M. 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.

[8] H.-Y. 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.

[9] W. Huang, M. Yi, X. Zhao, and Z. Jiang, "Towards the generalization of contrastive self-supervised learning," in *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[10] X. 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.

[11] X.-B. Nguyen *et al.*, "Self-supervised learning based on spatial awareness for medical image analysis," *IEEE Access*, vol. 8, pp. 162 973–162 981, 2020.

[12] Z. Zeng *et al.*, "Sese-net: Self-supervised deep learning for segmentation," *Pattern Recognition Letters*, vol. 128, pp. 23–29, 2019.

[13] O. Ciga *et al.*, "Self supervised contrastive learning for digital histopathology," *Machine Learning with Applications*, vol. 7, p. 100198, 2022.

[14] B. Jiménez *et al.*, "Comparison of the mayo endoscopy score and the ulcerative colitis endoscopy index of severity and the ulcerative colitis colonoscopy index of severity," *Endosc. Int. Open*, vol. 9, no. 2, pp. E130–E136, 2021.

[15] R. Erichsen *et al.*, "Increased risk of colorectal cancer development among patients with serrated polyps," *Gastroenterology*, vol. 150, no. 4, pp. 895–902.e5, Apr. 2016.

[16] I. Misra and L. v. d. 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.

[17] M. Gutmann and A. 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.

[18] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "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.

[19] X. 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. 12 586–12 595.

[20] Z. Xu, S. Ali, S. Gupta, S. Leedham, J. E. East, and J. Rittscher, "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.

[21] Z. Xu *et al.*, "Self-supervised approach for a fully assistive esophageal surveillance: Quality, anatomy and neoplasia guidance," in *Cancer Prevention Through Early Detection*, 2022, pp. 14–23.

[22] Z. 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.

[23] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[24] D. Jha *et al.*, "Kvasir-SEG: A segmented polyp dataset," in *International Conference on Multimedia Modeling*, 2020, pp. 451–462.

[25] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized medical imaging and graphics*, vol. 43, pp. 99–111, 2015.

[26] G. Polat *et al.*, "Labeled images for ulcerative colitis (LIMUC) dataset," 2022.

[27] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[29] R. W. Stidham *et al.*, "Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis," *JAMA network open*, vol. 2, no. 5, pp. e193 963–e193 963, 2019.

[30] M. F. Mokter, J. Oh, W. Tavanapong, J. Wong, and P. C. d. Groen, "Classification of ulcerative colitis severity in colonoscopy videos using vascular pattern detection," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2020, pp. 552–562.

[31] T. Ozawa *et al.*, "Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis," *Gastrointestinal endoscopy*, vol. 89, no. 2, pp. 416–421, 2019.

[32] B. G. Becker *et al.*, "Training and deploying a deep learning model for endoscopic severity grading in ulcerative colitis using multicenter clinical trial data," *Therapeutic advances in gastrointestinal endoscopy*, vol. 14, 2021.

[33] J. Y. Lee *et al.*, "Real-time detection of colon polyps during colonoscopy using deep learning: systematic validation with four independent datasets," *Scientific reports*, vol. 10, no. 1, pp. 1–9, 2020.

[34] X. Zhang *et al.*, "Real-time gastric polyp detection using convolutional neural networks," *PloS one*, vol. 14, no. 3, p. e0214133, 2019.

[35] H. A. 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.

[36] Y. Shin *et al.*, "Automatic colon polyp detection using region based deep CNN and post learning approaches," *IEEE Access*, vol. 6, pp. 40 950–40 962, 2018.

[37] Z. Zhou *et al.*, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[38] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "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.

[39] J. Wei, Y. Hu, R. Zhang, Z. Li, S. K. Zhou, and S. Cui, "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.

[40] A. Srivastava *et al.*, "Msrf-net: A multi-scale residual fusion network for biomedical image segmentation," *Journal of Biomedical and Health Informatics*, vol. 26, no. 5, pp. 2252–2263, 2021.

[41] M. Jaderberg *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[42] A. Sinha and J. Dolz, "Multi-scale self-guided attention for medical image segmentation," *IEEE journal of biomedical and health informatics*, vol. 25, no. 1, pp. 121–130, 2020.

[43] Q. Zhao *et al.*, "Adasan: Adaptive cosine similarity self-attention network for gastrointestinal endoscopy image classification," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 2021, pp. 1855–1859.

[44] R. Gu *et al.*, "Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," *IEEE transactions on medical imaging*, vol. 40, no. 2, pp. 699–711, 2020.

[45] T. Chen *et al.*, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[46] K. 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.

[47] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European conference on computer vision*. Springer, 2016, pp. 69–84.

[48] S. Ramesh, V. Srivastav, D. Alapatt, T. Yu, A. Murali, L. Sestini, C. I. Nwoye, I. Hamoud, S. Sharma, A. Fleurentin, G. Exarchakis, A. Karargyris, and N. Padoy, "Dissecting self-supervised learning methods for surgical computer vision," *Medical Image Analysis*, vol. 88, p. 102844, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841523001044

[49] M. Gutmann and A. 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.

[50] M. Misawa *et al.*, "Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video)," *Gastrointestinal endoscopy*, vol. 93, no. 4, pp. 960–967, 2021.

[51] C.-Y. Chuang *et al.*, "Debiased contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 8765–8775, 2020.

[52] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[53] T.-Y. Lin *et al.*, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[54] O. Ronneberger *et al.*, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[55] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, 2018.

[56] B. Dong, W. Wang, D.-P. Fan, J. Li, H. Fu, and L. Shao, "Polyp-pvt: Polyp segmentation with pyramid vision transformers," *arXiv preprint arXiv:2108.06932*, 2021.

[57] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*. PMLR, 2021, pp. 10 096–10 106.

[58] ——, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97, 09–15 Jun 2019, pp. 6105–6114.

[59] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[60] F. van der Sommen, J. de Groof, M. Struyvenberg, J. van der Putten, T. Boers, K. Fockens, E. J. Schoon, W. Curvers, P. de With, Y. Mori, M. Byrne, and J. J. G. H. M. Bergman, "Machine learning in GI endoscopy: practical guidance in how to interpret a novel field," *Gut*, vol. 69, no. 11, pp. 2035–2045, Nov. 2020.

[61] S. Ali *et al.*, "Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge," *Sci Rep*, vol. 14, p. 2032, 2024.