Contrastive Self-Supervised Learning on Crohn's Disease Detection

Jing Xing, Harold Mouchère

Contrastive Self-Supervised Learning on Crohn's Disease Detection

1st Jing XING
South China University of Technology
GuangZhou, China
Nantes Université, École Centrale Nantes, CNRS
LS2N, UMR 6004, F-44000 Nantes, France
xjjsdsb@gmail.com

2nd Harold Mouchère
Nantes Université, École Centrale Nantes, CNRS
LS2N, UMR 6004, F-44000 Nantes, France
harold.mouchere@univ-nantes.fr

***Abstract*—Crohn's disease is a type of inflammatory bowel illness that is typically identified via computer-aided diagnosis (CAD), which employs images from wireless capsule endoscopy (WCE). While deep learning has recently made significant advancements in Crohn's disease detection, its performance is still constrained by limited labeled data. We suggest using contrastive self-supervised learning methods to address these difficulties which was barely used in detection of Crohn's disease. Besides, we discovered that, unlike supervised learning, it is difficult to monitor contrastive self-supervised pretraining process in real time. So we propose a method for evaluating the model during contrastive pretraining (EDCP) based on the Euclidean distance of the sample representation, so that the model can be monitored during pretraining. Our comprehensive experiment results show that with contrastive self-supervised learning, better results in Crohn's disease detection can be obtained. EDCP has also been shown to reflect the model's training progress. Furthermore, we discovered some intriguing issues with using contrastive self-supervised learning for small dataset tasks in our experiments that merit further investigation.**

***Index Terms*—Deep Learning, Self-Supervised Learning, Contrastive Learning, Crohn's Disease, Medical images classification**

I. INTRODUCTION

Crohn's Disease is a common bowel disease [1]. Wireless Capsule Endoscopy (WCE) is commonly used by doctors to aid in diagnosis. However, counting and identifying lesions of Crohn's disease in WCE videos is time-consuming. Many methods [2] have been developed to automatically detect intestinal abnormalities in WCE images, such as ulcer, bleeding, and erosion. These methods can be divided into two main categories. One is using Support Vector Machine(SVM) [3]–[14] with manually designed features. The other is based on deep learning [15]–[20]. In the early years, researchers primarily used SVM as classifiers based on manually designed features such as local binary pattern [21], scale invariant feature transform [21], texture feature [22], color feature [23], etc. However, deep learning has demonstrated tremendous power in a number of fields in recent years. The accuracy of Crohn's disease detection has increased significantly with deep neural networks to over 90% [2].

978-1-6654-6819-0/22/ $31.00 ©2022 IEEE

As shown in Fig. 1, the images for Crohn's disease detection are taken from WCE video, thus there are numerous unlabeled images. Nevertheless, because labeling WCE image is very time-consuming and requires professional doctors, labelled data is scarce. As a result, deep learning's effectiveness in the field of medical images, including Crohn's disease detection, is massively diminished. In this situation, self-supervised learning is a viable option. Given that it can train the network with unlabeled data to make up for the lack of enough labeled data. Self-supervised methods based on contrastive learning have recently been proven to be more effective in image classification tasks. The CrohnIPI [19] dataset that we use is a professional public dataset that contains well-labeled data. To the best of our knowledge, existing research works use only self-supervised learning for Crohn's disease detection. Besides, most self-supervised studies use very large datasets as benchmark, such as ImageNet [24]. However, WCE images are usually small and have few categories which is very different to ImageNet. Considering the gap between ImageNet and WCE images, we compare several state-of-the-art contrastive self-supervised learning methods' performance on WCE data to find out that whether self-supervised learning can improve the performance of Crohn's disease detection.

During our experiments, we discovered that self-supervised training, also referred pretraining, unlike supervised training, cannot use the validation set to observe the pretraining process. This also makes selecting model of pretraining and conducting early-stopping impossible. We can only use finetuning to verify the model's performance after pretraining, which significantly prolongs the model's tuning process and makes selecting the best model during the pretraining process for downstream tasks difficult. Therefore, we propose *evaluation during contrastive pretraining* (EDCP), a low-cost method for real-time monitoring of pretraining progress. Based on the Euclidean distance between sample representations, this method computes a value that reflects how well network is trained. We compare the EDCP to the results of few-epoch finetuning and discovered that the EDCP results are consistent with the few-epoch finetuning results, which demonstrats that EDCP can be used effectively to monitor contrastive self-supervised learning pretraining.

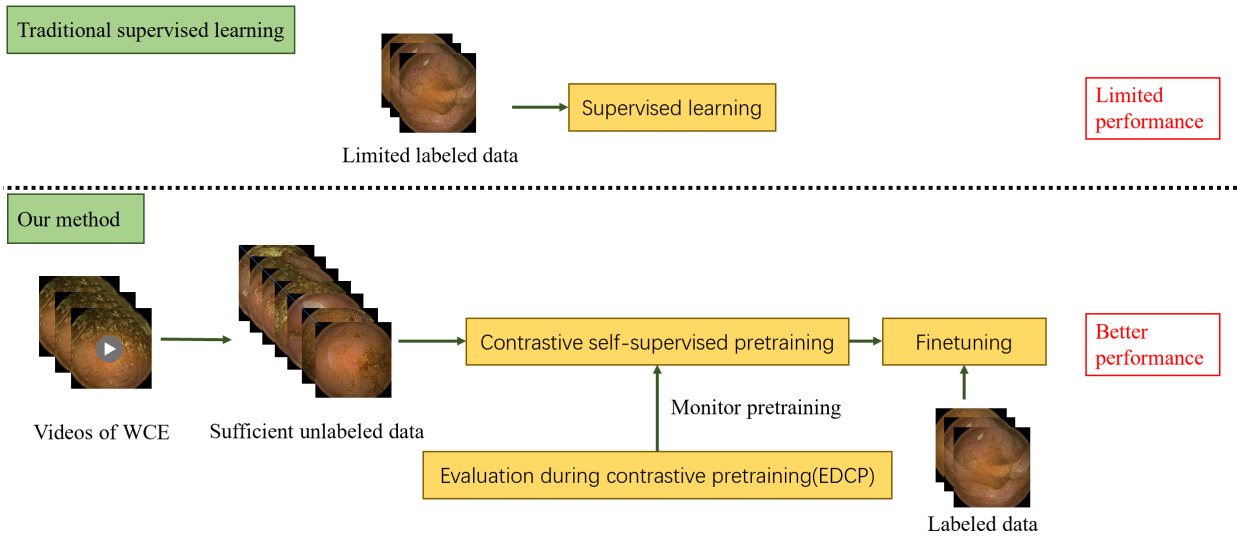

Fig. 1. Traditional supervised learning (top), our contrastive self-supervised learning method (bottom). Evaluation During Contrastive Pretraining(EDCP) is used to monitor pretraining(Bottom). The wireless capsule endoscopy (WCE) images are from CrohnIPI [19].

Furthermore, we observed some contrastive learning techniques that are useful on large datasets but are ineffective or even harmful for Crohn's disease detection in experiments. A larger batch size, for example, will generally produce better results [25], but it has no effect on Crohn's disease detection, and even too many negative sample pairs will degrade the model's performance. These anomalous phenomena merit further investigation, as they may improve the model's performance and have some reference significance for similar classification tasks of small dataset.

The following are our primary contributions: **(i)**We are the first to apply self-supervised learning to the detection of Crohn's disease in images. **(ii)** We outperform traditional supervised learning in the detection of Crohn's disease. **(iii)** We propose using evaluation during contrastive pretraining (EDCP) to track the progress of contrastive self-supervised training.

II. PRELIMINARY AND RELATED WORK

Crohn's Disease Detection. Computer-aided diagnosis (CAD) has become an active search area in the past few years. WCE is crucial for the diagnosis of Crohn's disease because it allows medical professionals to see inside the patient's bowel in order to identify Crohn's disease as soon as possible. However, diagnosing with WCE images takes a long time for medical experts. Many methods for automatically detecting diseases in WCE based on deep learning have recently been proposed, and excellent results have been obtained [2]. Deep learning methods based on convolutional neural networks (CNN) have been proven to outperform prior machine learning methods which based on support vector machines (SVM) and manually designed features. However, there is still room for improvement due to the lack of labeled data. In fact, insufficient labeled data is a common issue in the medical imaging field. For example, Seguí et al [26] found that 10-fold cross-validation increase in training data size improved the accuracy by 3% for motility movement classification in WCE. WCE images have the significant advantage of being easy to obtain a large number of unlabeled images since tens of thousands of frames are contained in each WCE video. To make use of these unlabeled data, we employ self-supervised learning.

Contrastive Self-Supervised Learning. Self-supervised learning (SSL) has recently achieved great success in a variety of fields. Liu et al. [27] summarize the mainstream self-supervised learning into three broad categories: generative, contrastive, and generative-contrastive. In generative tasks such as image colorization [28] and inpainting [29], generative and generative-contrastive learning are particularly effective. Contrastive learning, on the other hand, outperforms in classification tasks, particularly in computer versions (CV). Among contrastive learning methods, instance-instance contrast is proved perform better than context-instance contrast [27]. Many instance-instance [25], [30]–[35] have performed admirably on various CV benchmarks. The object of contrast in contrastive learning is commonly referred to as negative samples and positive samples. The different embeddings produced by transforming one image are referred to as positive samples from each other, whereas the embeddings produced by transforming different images are referred to as negative samples from each other. To obtain positive and negative samples, the contrastive learning network typically has two similar branches, and is therefore also known as siamese network [36]. The objective of contrastive learning is to maximize the similarity of negative pairs while minimize the similarity of positive pairs. InfoNCE [37] loss is commonly used to achieve this goal. The InfoNCE loss for sample q with distance measured by dot product is:

$$\mathcal{L}_q = -\log \frac{\exp\left(q \cdot k_+/\tau\right)}{\sum_{i=0}^{K} \exp\left(q \cdot k_i/\tau\right)}, \tag{1}$$

where τ is a temperature hyper-parameter. The sum is over one positive sample k_+ and K negative samples. Many studies [25], [30], [33] have shown that having a sufficient number of negative samples is critical for model performance and avoiding clloaps during pretraining.

State of the Art (SOTA) Methods. In the field of contrastive self-supervised learning, method iteration has been remarkably rapid in recent years. InvaSpread [38] presents an end-to-end method for contrastive learning. Following InvaSpread, SimCLR made significant progress through refining the experiment and network structure including adding MLP head, more epochs of pretraining, much bigger batch size and more complex data augmentations. These manipulations are straightforward and effective, and they are followed by methods proposed later. Meanwhile, MoCo [30] introduced momentum encoder to reduce inconsistency caused by rapidly changing encoder, based on InstDisc [39], which proposed memory bank that use a small amount of computational resources while producing a large number of negative samples. MoCo v2 later combined the useful SimCLR strategies on the basis of moco to achieve better results. Positive and negative pairs are sample-based in MoCo v2 and SimCLR, whereas in Barlow Twins [33], they are based on filters that generate embeddings. They believe that this will allow filters to learn more different features and reduce the redundancy of information. Barlow Twins does not require a large batch size or memory bank, but it does require more channels in the MLP layer in order to obtain enough negative samples. Most contrastive self-supervised learning methods proposed prior to BYOL [34] require negative samples to avoid trivial solutions. For example, if the network only needs to reduce the similarity of positive pairs, regardless of the input, the network can output the same feature vector. BYOL produces excellent results with only positive sample pairs and its performance is robust to small batch sizes. Soon after, SimSiam [36] proposed a simplified version of BYOL which remove momentum encoder and the predictor head while slightly reducing performance.

III. Methods

A. Baseline

Our aim is to find out whether contrastive self-supervised learning can improve the performance of Crohn's disease detection. For comparison, the traditional supervised learning method is used. Following [19], the network structure of baseline consists of an encoder and a fully connected layer. The encoder is ResNet [40] without final fully connected layer.

B. Contrastive Self-Supervised Methods Comparison

SOTA methods mentioned in II can be divided into three categories according to the factors contrasted: positive and negative pairs based on instances(MoCo, MoCo v2, SimCLR, etc.), positive and negative pairs based on filters(Barlow Twins), and only positive pairs based on instances(BYOL, SimSiam). The best methods from each of these three categories for detecting Crohn's disease detection will be chosen.

TABLE I
Top-1 and top-5 accuracies (in %) under linear evaluation on ImageNet. All models use ResNet-50 as encoder. The batch size is for pretraining. These results are come from corresponding papers.

Method	Top-1	Top-5	Batch Size
Supervised	76.5	-	-
MoCo	60.6	-	256
SimCLR	69.3	89.0	4096
MoCo v2	71.1	-	256
SimSiam	71.3	-	256
Barlow Twins	73.2	91.0	4096
BYOL	74.3	91.6	4096

We are primarily concerned with two aspects when comparing methods: performance and robustness to batch size. A general comparison of the performance of self-supervised learning is to compare the performance of linear evaluation on ImageNet which is shown in Table I. According to the Table I, it can be seen that MoCo v2, SimSiam, Barlow Twins, and BYOL perform well. Barlow Twins and BYOL stand out among them. When comparing the robustness of methods to batch size, the study [33] in Barlow Twins shows that the performance of Barlow Twins and BYOL are still good when the batch size decreases, while the performance of SimCLR drop a lot. Because the loss strategy in the BYOL and Barlow twins determines that their performance is less dependent on a large number of negative samples. MoCo v2's performance is also robust to batch size since its negative samples are primarily drawn from memory bank which is not influenced by batch size. As a result, three methods were chosen from the three categories based on their good performance and robustness of batch size: MoCo v2, BYOL, and Barlow Twins. Since the hyperparameters of each method are carefully tuned, we will try to keep the original hyperparameters of these methods during the experiment for fairness.

C. Framework

The three methods have similar structure which shown in the Fig. 2. The input image is augmented t and t' to produce two distinct views for two branches. The network's goal is that the representations generated by two branches for two views of the same image are similar, while the representations generated for different input images are dissimilar. The network, with the exception of the BYOL predictor, is made up of two symmetrical branches. Each branch is made up of an encoder f, which is a CNN followed by an projector network MLP. The encoder structure of the three networks is ResNet without final full-connected layer, but the projector structure is different. The projector MLP structures of MoCo v2, BYOL, and Barlow Twins are 512-128, 4096-256, and 8192-8192-8192, respectively. Following SimCLR, the projector MoCo v2 consists in a linear layer with output size 512 follwed by rectified linear units (ReLU) and a final linear layer with output dimension 256. The projector of BYOL consists in a linear

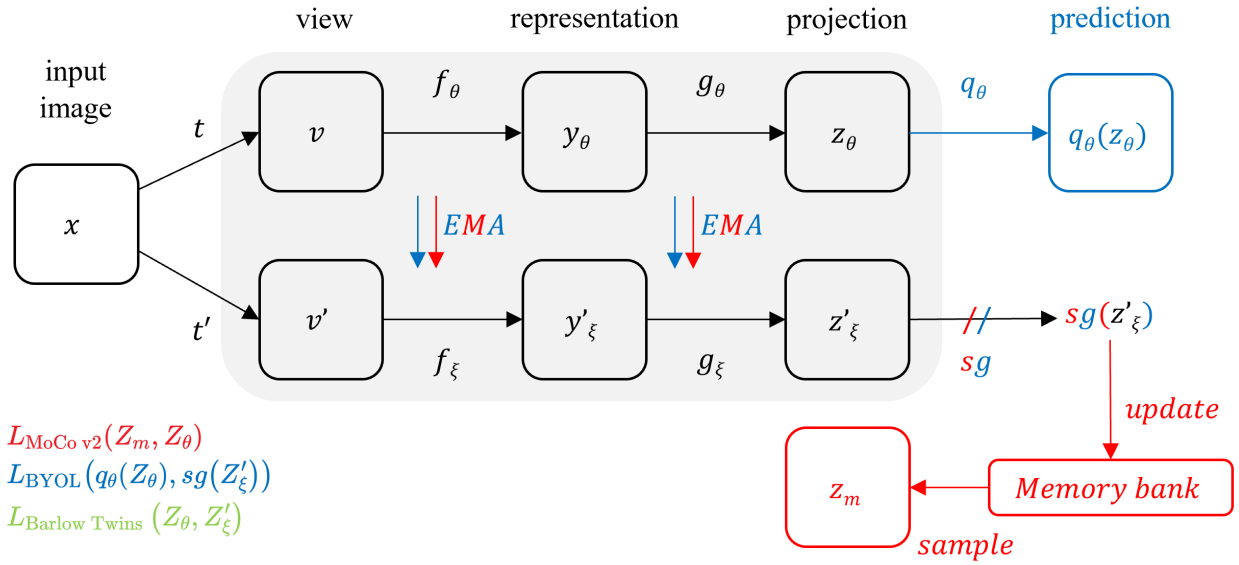

Fig. 2. The architecture of MoCo v2, BYOL and Barlow Twins. The black part represents the structure shared by all three methods, the red part represents MoCo v2, and the blue part represents BYOL. θ are the trained weights, ξ are exponential moving average of θ and sg denotes stop-gradient. At the end of training, everything except f_θ is discarded, and y_θ is used as the image representation. The embeddings used for calculating loss by three methods are shown in the lower left corner.

layer with output size 4096 followed by batch normalization, ReLU, and a final linear layer with output dimension 256. The predictor of BYOL uses the same architecture as projector. The projector of Barlow Twins has three linear layers, each with 8192 output units. A batch normalization layer and ReLU are placed after the projector's first two layers. We can see that Barlow Twins require large output dimension to provide a large number of negative samples in order to perform well. Besides, despite the simplicity of MoCo v2's MLP structure, MoCo v2 requires memory banks to provide a large number of embeddings as negative samples.

The parameters of Barlow Twins are all updated by back-propagation. However, in the network structure of BYOL and MoCo v2, the stop gradient operation is performed at the end of one branch, and the parameter ξ of this branch is updated by exponential moving average (EMA) strategy which formulated by (2).Back-propagation updates the parameter θ of another branch.

$$\xi \leftarrow \tau\xi + (1-\tau)\theta \tag{2}$$

where $\tau \subset [0, 1)$ is a coefficient.

The embeddings used by the three methods to calculate the loss are different, as shown in the Fig. 2. MoCo v2 employs InfoNCE loss between projection z_θ and embedding z_m sampled from the memory bank. The loss in BYOL is calculated as the mean squared error between the $l2$-normalized projection $sg(z'_\xi)$ and the prediction $q_\theta(z_\theta)$. After normalize the embeddings along the batch dimension, Barlow Twins cleverly transposes one projection z_θ and multiplies it with another projection z'_ξ to produce a cross-correlation matrix with diagonal and off-diagonal elements representing the cosine similarity of positive and negative sample pairs, respectively. The identity matrix is used as the label, and the loss is calculated as the mean square error of the cross-correlation matrix and the identity matrix in Barlow Twins.

D. Evaluation During Contrastive Petraining (EDCP)

The validation set is commonly used in supervised training to monitor training progress in order to avoid overfitting or insufficient training, and it can also be used to select models. However, because the evaluation of self-supervised learning is finetuning when pretraining is finished, evaluating the network during pretraining is difficult. Finetuning after each epoch during pretraining is a simple and straightforward method, but it adds a significant amount of time and computational resource consumption. Another straightforward approach is to monitor pretraining process directly through observing the loss, but these losses are not related to the classification task. First, this is an indirect method. There is inconsistency because the downstream task uses only the encoder rather than the entire pretraining network. Second, the loss may gradually decrease with training due to the strategy of some methods. During pretraining, for example, the momentum of BYOL will gradually increase, resulting in closer parameters of the two branches of the network, which brings the generated representation closer to each other and eventually leads to a smaller loss.

To address this problem, we design a method for real-time monitoring of model training progress that can be quickly computed and widely used in any contrastive learning pretraining process, called evaluation during contrastive pretraining (EDCP).

Based on the goal of contrastive learning, which is to decrease the positive sample pair's feature similarity close and increase the negative sample pair's feature similarity. Euclidean distance is used to measure similarity. We input

a small amount of labeled data into the encoder to get the corresponding representations, and then calculate the Euclidean distance of the positive and negative sample pairs. Two subsets are sampled with no overlap, each subset has the same number of samples with different labels, and they are converted into representations set M, N by the encoder. Then subtract the distance of the negative representation pairs from the distance of the positive sample pair to get S_{EDCP}. For example, for a representation m_l from M with label l, find the closest representation $m_{l'}$ in M with different label l' and the closest representation n_l, n'_l from N. Then use (5) to calculate the $L2$ distance of m_l and other representations to get $d(m_l, m'_l)$,$d(m_l, n_l)$ and $d(m_l, n'_l)$. Then subtract $d(m_l, n'_l)$ and $d(m_l, m'_l)$ by $d(m_l, n_l)$ to get the distance value for m_l. Sum and all distance for each representation in M, N will get S_{EDCP}. The calculation process is as follows:

$$S_{EDCP}(M, N) = \frac{1}{K}[D(M, N) + D(N, M)] \tag{3}$$

$$D(M, N) = \sum_i [\min_j d(m_{i,l}, n_{j,l}) - \frac{1}{2}(\min_j d(m_{i,l}, n_{j,l'}) + \min_j d(m_{i,l}, m_{j,l'}))] \tag{4}$$

$$d(m, n) = \|m - n\|_2 \tag{5}$$

Where $m_{l,i}$ is i-th sample of M with label l, $n_{j,l}$ is j-th sample of N with label l, K is the sum number of M, N.

The smaller the value of S_{EDCP}, the better the encoder is trained. So that monitoring process of pretraining can be achieved by observing the curve of S_{EDCP}. And the encoder at the point where S_{EDCP} reaches a minimum is likely to be the best pretrained encoder. Meanwhile, to assess if the trend of curve of S_{EDCP} is correct, we conduct one-epoch-finetuning during pretraining. Because Crohn's disease detection is a relatively simple task that requires only a small number of epochs to complete finetuning, one-epoch-finetuning can, to some extent, represent the final result of finetuning at a lower cost. Therefore, the ability of EDCP to reflect the encoder's pretraining trend can be determined by comparing the results of EDCP and one-epoch-finetuning.

IV. Experiment

To explore whether contrastive self-supervised learning can achieve better performance than traditional supervised learning on Crohn's disease detection, we first performed a baseline experiment to obtain results for comparison. Then, for each of the three contrastive self-supervised learning methods, perform pretraining and finetuning. The finetuning and baseline settings are the same. EDCP and one-epoch-finetuning are used during pretraining. All of our experiments were conducted on a 16GB Tesla T4.

A. Dataset.

For baseline and finetuning experiments, 3484 labeled images of CrohnIPI [19] are used, including 1360 pathological images and 2124 non-pathological images. For pretraining, 35053 unlabeled images were obtained from two WCE videos, different from the labeled images. It is worth noting that the majority of the datasets used in the WCE image detection task research are private datasets or unspecified subsets of public databases. However, CrohnIPI is a publicly available and well-labeled dataset that can be widely used as a benchmark for Crohn's disease detection.

B. Baseline & Finetuning

Image Augmentations. The input image will be resized to 256x256 for training after a random horizontal flip, vertical flip, and rotation. During testing, the images are resized to 256x256.

Optimization. Following the settings in CrohnIPI [19], ResNet is used as the encoder, followed by a fully connected layer as the classification head. For back-propagation, we use the cross-entropy loss. Adam optimizer is used to optimize the cross-entropy loss. For the entire network, we use a learning rate of 0.0003 and train for 300 epochs. We use a batch size of 16 and 5-fold cross-validation. For data splitting, there is 20% labeled data for testing and the rest for training and validation. The ratio of the images used for training and validation is 1:4. We use validation set to validate the model after each training epoch and select the model with the highest validation accuracy for test.

C. Pretraining

For fair comparison, we use the same encoder, number of epochs and batch size when pretraining the three contrastive learning methods, which are resnet18 (without the final classification layer), 300, and 96, respectively

Image Augmentation. The data augmentation of pretraining comes after the augmentation in SimCLR [25]. First, a random patch of the image is chosen and resized to 256x256 with a random horizontal flip, then color distortion is applied with a random sequence of brightness, contrast, saturation, hue, and optional grayscale conversion adjustments. Gaussian blur and solarization are applied to the patch as a final step.

MoCo v2 on CrohnIPI. For optimization, MoCo v2 uses SGD as optimizer. The momentum of SGD is 0.9, and the weight decay of SGD is 0.0001. At 120 and 160 epochs, the learning rate was multiplied by 0.1. The learning rate is set to 0.03. A learning rate schedule based on cosine is used. The memory bank size is changed to 65472 instead of 65536 to accommodate the batch size of 96. Because the size of memory bank must be an integer multiple of the batch size

BYOL on CrohnIPI. BYOL uses LARS Optimizer [41] with a cosine decay learning rate schedule and 10 warm-up epochs. The baseline learning rate was set to 0.2 and was linearly scaled [42] with batch size ($LearningRate = 0.2 \times BatchSize/256$). BYOL uses a global weight decay parameter of $1.5 \cdot 10^{-6}$ while omitting the biases and batch normalization parameters from both LARS adaptation and weight decay. The target network's exponential moving average parameter τ starts at 0.996 and is raised to one

TABLE II
TEST ACCURACY OF BASELINE WITH DIFFERENT ENCODERS.

encoder	ResNet-18	ResNet-34	ResNet-50
Test accuracy(%)	93.8	93.6	93.1

during training. BYOL specifically set k the current training step and K the maximum number of training steps to, $\tau \triangleq 1-(1-\tau_{\text{base}}) \cdot(\cos(\pi k/K)+1)/2$.

Barlow Twins on CrohnIPI. Barlow Twins follow the optimization procedure outlined in BYOL [34] which use LARS optimizer for optimization. They utilize learning rates of 0.0048 for the biases and batch normalization parameters and 0.2 for the weights. They divide the learning rate by 256 and then multiply the result by the batch size. After a 10-epoch warm-up period, they employ a cosine decay schedule to reduce the learning rate by a factor of 1000 [43]. The trade-off parameter λ is set to $5 \cdot 10^{-3}$. A weightdecay parameter of $1.5 \cdot 10^{6}$ is used. The biases and batch normalization parameters are excluded from LARS adaptation and weight decay.

D. Evaluation During Contrastive Pretraining

From the labeled dataset, we extracted two subsets. Each subset has 100 total images, of which 50 are pathological and 50 are non-pathological. We feed the image into the encoder every 20 epochs of pretraining to generate the corresponding representation. The S_{EDCP} is calculated using (3). At the same time, we perform one-epoch-finetuning every 20 epochs and get value of accuracy. Theoretically, accuracy rises as S_{EDCP} increases and vice versa. By comparing the trend of S_{EDCP} and the trend of accuracy of one-epoch-finetuning, it is possible to demonstrate whether the EDCP method can be used to monitor the process of contrastive self-supervised learning pretraining.

V. EXPERIMENT RESULTS

A. Baseline

We have tried use Resnet18, Resnet34 and Resnet50 and find that the network with Resnet18 as encoder get best performance as shown in Table II. And the curve of validation and training accuracy shown in Fig. 3 demonstrates that 50 epochs are sufficient for the network to converge. Since detecting Crohn's disease is not a difficult task and the labeled dataset is small, the model can converge and achieve good accuracy without training for a long time.

B. Contrastive Self-Supervised Training

We utilize ResNet-34 as the encoder and pretrain MoCo v2. BYOL and Barlow Twins for 300 epochs with a batch size of 96. After pretraining, we finetuning with encoder initialized by pretrained parameters for 50 epochs. The results of finetuning are shown in Table III.

It can be seen from Table III that BYOL gets best performance in three self-supervised methods. It is intriguingly to find that MoCo v2 get relative lower performance. The reason may be due to the large number of positive samples considered as negative samples in binary classification tasks. This issue will be further investigated in VI-A.

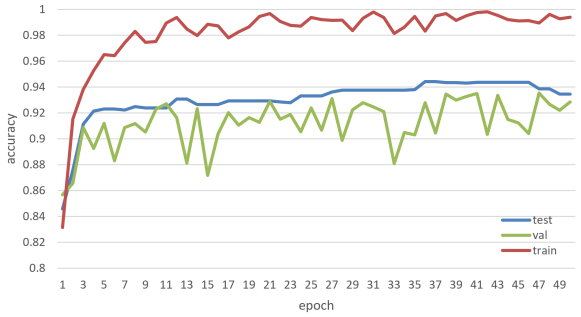

Fig. 3. Training, validation and test accuracy of baseline (with ResNet-18 encoder) to the number of training epochs in one of 5-fold cross-validation.

TABLE III
FINETUNING ACCURACY FOR MOCO V2 BYOL AND BARLOW TWINS. THE ENCODER IS RESNET35.

Methods	Accuracy (%)
MoCo v2	90.6
BYOL	93.2
Barlow Twins	91.5

According to studies on contrastive self-supervised learning, the batch size, encoder size, and number of pretraining epochs could all have a significant impact on performance. According to the result of baseline experiment, smaller encoder could have better performance. Therefore, based on BYOL, we compare the performance with different encoder, epochs and batch size. The results are shown in Table IV which also includes the baseline results for comparison.

From Table IV, about the size of encoder we can see that the performance is in line with baseline, that the smaller encoder performs better. Unlike in the contrastive self-supervised learning research on ImageNet, the larger encoder gets better performance does not work here. Furthermore, we discover that more epochs in BYOL pretraining can improve encoder training while larger batch size does not. When the ResNet-18 is pretrained after 600 epochs with 256 batch size, its performance outperforms the baseline which is inspiring. This result prove that contrastive self-supervised learning can improve the performance of Crohn's disease detection. However, there is still a lot of room for improvement. Since our primary goal is to investigate whether contrastive self-supervised learning is effective for Crohn's disease detection and which kind of contrastive method is more effective, we haven't done much tuning of the network structure and training parameters. We believe that many operations, such as training for longer epochs, using smaller encoders and simpler MLP layers, and using more specifically aimed data augmentation for WCE images could improve performance.

C. Evaluation During Contrastive Pretraining

When the encoder is better pretrained, the accuracy of one-epoch-finetuning should increase, while the representation

TABLE IV
OPTIMIZATION OF EPOCHS, BATCHS SIZE AND SIZE OF ENCODER OF PRETRAINING. WHERE ACCURACY EXCEEDS THE BASELINE, IT IS HIGHLIGHTED IN RED. BYOL IS USED AS PRETRAINING METHOD.

encoder	number of epoch	batch size	accuracy(%)
Baseline_ResNet-18	-	-	93.8
Baseline_ResNet-34	-	-	93.5
ResNet-34	300	96	93.2
ResNet-18	300	96	93.6
ResNet-18	600	256	94.0
ResNet-18	300	256	93.6

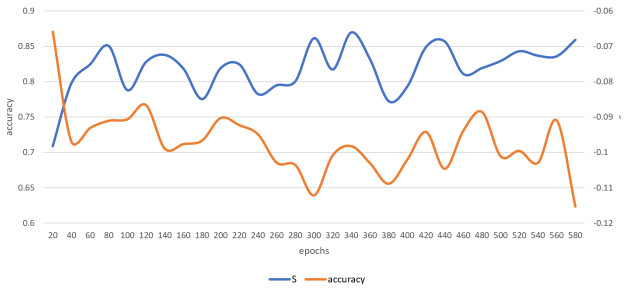

Fig. 4. The curve of one-epoch-finetuning accuracy and distance S of ECDP during pretraining. S is S_{EDCP} in (3).

distance of positive samples decreases and the distance between negative sample pairs increases, resulting in a decrease in s_{EDCP}. The results of EDCP and one-epoch-finetuning accuracy are shown in Fig. 4. The curves of Fig. 4 shows that the trends in accuracy and s_{EDCP} are roughly opposite. When the training reaches 300 epoch, for example, the accuracy and s_{EDCP} reach a local maximum and a local minimum, respectively. Besides, from the overall downward trend of s_{EDCP}, we can also know that the performance increases with the increase of the number of epochs, and has not converged, which is consistent with the results shown in the Table IV. As a result, the curve demonstrates that s_{EDCP} can effectively reflect the training degree during the pretraining process and can be used as an early-stopping and encoder selection tool in pretraining.

VI. DISSCUSSION

A. Noisy Negative Samples of MoCo v2

According to study in MoCo [30], increasing the number of features taken from the memory bank improves performance, similar to increasing the batch size in SimCLR. However, in our experiment, MoCo v2 does not perform well when there are a large number of negative sample pairs, according to the results shown in V-B. We believe that it is due to the noisy sample pairs, in which many positive samples are misidentified as negative samples. Because MoCo is an instance-instance contrastive self-supervised learning method. During pretraining, all other samples are negative samples for a sample. This strategy, however, is harmful for binary classification tasks. Because there will be a large number of samples that are considered negative during pretraining but positive during finetuning. This causes the encoder to undergo conflicting training and it becomes more severe as the number of negative samples increases when pretraining. There is no such issue in BYOL or Barlow Twins. Because BYOL does not use negative samples, and the negative samples used by the Barlow Twins are based on filters. The results of Table V show that reducing the number of negative samples improves the performance of MoCo v2 on Crohn's detection significantly. As a result, the sample-based contrastive self-supervised method may not be appropriate for binary classification or tasks on small dataset. Furthermore, the noise cancellation for samples merits further investigation.

TABLE V
MEMORY BANK COMPARISON

Memory bank	accuracy(%)
576	91.5
65536	90.6

B. Self-Supervised Learning on Small Dataset

Research on self-supervised methods is generally conducted on large benchmarks, such as ImageNet, because this provides a more objective evaluation of models and methods. Nevertheless, in the field of medical images, many datasets are small and the data diversity is limited due to the focus on specific disease or body part, such CrohnIPI. As a result, some methods and techniques that work on ImageNet may not work on small datasets. For instance, while a larger encoder and batch size are typically effective in improving performance on ImageNet, they are ineffective and even worsen performance in our experiments. Network and training strategy design for small datasets is lacking. These methods are critical in the practical application of deep learning methods, such as classification and segmentation of various medical image tasks. We hope that this study inspires more research into self-supervised learning on small datasets and we will also continue to investigate the application of self-supervision on small data sets.

VII. CONCLUSION

To alleviate the lack of sufficient labeled data for Crohn's disease detection, we investigated the use of contrastive self-supervised learning methods combined with unlabeled data, which improved the accuracy to 94%. We also propose an evaluation during contrastive pretraining (EDCP) method for monitoring pretraining and model selection for self-supervised pretraining. Finally, we observed some intriguing issues with using contrastive self-supervised learning for small dataset tasks in our experiments that merit further investigation.

REFERENCES

[1] R. Eliakim, "The impact of panenteric capsule endoscopy on the management of crohn's disease," *Therapeutic Advances in Gastroenterology*, vol. 10, no. 9, pp. 737–744, 2017.

[2] S. Soffer, E. Klang, O. Shimon, N. Nachmias, R. Eliakim, S. Ben-Horin, U. Kopylov, and Y. Barash, "Deep learning for wireless capsule endoscopy: a systematic review and meta-analysis," *Gastrointestinal endoscopy*, vol. 92, no. 4, pp. 831–839, 2020.

[3] O. Hajo-Maghsoudi, A. Talebpour, H. Soltanian-Zadeh, and N. Haji-Maghsoodi, "Segmentation of crohn, lymphangiectasia, xanthoma, lymphoid hyperplasia and stenosis diseases in wce," in *Quality of Life through Quality of Information*. IOS Press, 2012, pp. 143–147.

[4] Y. Chen and J. Lee, "Ulcer detection in wireless capsule endoscopy video," in *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 1181–1184.

[5] V. S. Charisis, L. J. Hadjileontiadis, C. N. Liatsos, C. C. Mavrogiannis, and G. D. Sergiadis, "Capsule endoscopy image analysis using texture information from various colour models," *Computer methods and programs in biomedicine*, vol. 107, no. 1, pp. 61–74, 2012.

[6] A. Eid, V. S. Charisis, L. J. Hadjileontiadis, and G. D. Sergiadis, "A curvelet-based lacunarity approach for ulcer detection from wireless capsule endoscopy images," in *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*. IEEE, 2013, pp. 273–278.

[7] P. Szczypiński, A. Klepaczko, M. Pazurek, and P. Daniel, "Texture and color based image segmentation and pathology detection in capsule endoscopy videos," *Computer methods and programs in biomedicine*, vol. 113, no. 1, pp. 396–411, 2014.

[8] J.-Y. Yeh, T.-H. Wu, W.-J. Tsai *et al.*, "Bleeding and ulcer detection using wireless capsule endoscopy images," *Journal of Software Engineering and Applications*, vol. 7, no. 05, p. 422, 2014.

[9] D. K. Iakovidis and A. Koulaouzidis, "Automatic lesion detection in capsule endoscopy based on color saliency: closer to an essential adjunct for reviewing software," *Gastrointestinal endoscopy*, vol. 80, no. 5, pp. 877–883, 2014.

[10] Y. Yuan, J. Wang, B. Li, and M. Q.-H. Meng, "Saliency based ulcer detection for wireless capsule endoscopy diagnosis," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 2046–2057, 2015.

[11] V. S. Charisis and L. J. Hadjileontiadis, "Use of adaptive hybrid filtering process in crohn's disease lesion detection from real capsule endoscopy videos," *Healthcare technology letters*, vol. 3, no. 1, pp. 27–33, 2016.

[12] A. Liaqat, M. A. Khan, J. H. Shah, M. Sharif, M. Yasmin, and S. L. Fernandes, "Automated ulcer and bleeding classification from wce images using multiple features fusion and selection," *Journal of Mechanics in Medicine and Biology*, vol. 18, no. 04, p. 1850038, 2018.

[13] S. Alotaibi, S. Qasim, O. Bchir, and M. M. Ben Ismail, "Empirical comparison of visual descriptors for multiple bleeding spots recognition in wireless capsule endoscopy video," in *International Conference on Computer Analysis of Images and Patterns*. Springer, 2013, pp. 402–407.

[14] M. Souaidi and M. E. Ansari, "Multi-scale analysis of ulcer disease detection from wce images," *IET Image Processing*, vol. 13, no. 12, pp. 2233–2244, 2019.

[15] S. V. Georgakopoulos, D. K. Iakovidis, M. Vasilakakis, V. P. Plagianakos, and A. Koulaouzidis, "Weakly-supervised convolutional learning for detection of inflammatory gastrointestinal lesions," in *2016 IEEE international conference on imaging systems and techniques (IST)*. IEEE, 2016, pp. 510–514.

[16] S. Fan, L. Xu, Y. Fan, K. Wei, and L. Li, "Computer-aided detection of small intestinal ulcer and erosion in wireless capsule endoscopy images," *Physics in Medicine & Biology*, vol. 63, no. 16, p. 165001, 2018.

[17] T. Aoki, A. Yamada, K. Aoyama, H. Saito, A. Tsuboi, A. Nakada, R. Niikura, M. Fujishiro, S. Oka, S. Ishihara *et al.*, "Automatic detection of erosions and ulcerations in wireless capsule endoscopy images based on a deep convolutional neural network," *Gastrointestinal endoscopy*, vol. 89, no. 2, pp. 357–363, 2019.

[18] H. Alaskar, A. Hussain, N. Al-Aseem, P. Liatsis, and D. Al-Jumeily, "Application of convolutional neural networks for automated ulcer detection in wireless capsule endoscopy images," *Sensors*, vol. 19, no. 6, p. 1265, 2019.

[19] R. Vallée, A. de Maissin, A. Coutrot, H. Mouchère, A. Bourreille, and N. Normand, "CrohnIPI: An endoscopic image database for the evaluation of automatic Crohn's disease lesions recognition algorithms," in *SPIE Medical Imaging*, ser. Proc. SPIE, Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging, vol. 11317. Houston, France: SPIE, Feb. 2020, p. 61. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02518263

[20] A. de Maissin, R. Vallée, M. Flamant, M. Fondain-Bossiere, C. Le Berre, A. Coutrot, N. Normand, H. Mouchère, S. Coudol, C. Trang *et al.*, "Multi-expert annotation of crohn's disease images of the small bowel for automatic detection using a convolutional recurrent attention neural network," *Endoscopy International Open*, vol. 9, no. 07, pp. E1136–E1144, 2021.

[21] L. Yu, P. C. Yuen, and J. Lai, "Ulcer detection in wireless capsule endoscopy images," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012, pp. 45–48.

[22] Y. Che, *MM'12 : proceedings of the 20th ACM International Conference on Multimedia : October 29 - November 2, 2012, Nara, Japan*. ACM, 2012.

[23] B. Li and M. Q. Meng, "Computer-based detection of bleeding and ulcer in wireless capsule endoscopy images by chromaticity moments," *Computers in Biology and Medicine*, vol. 39, pp. 141–147, 2 2009, 3layerMLP+color feature.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[25] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020. [Online]. Available: https://arxiv.org/abs/2002.05709

[26] S. Seguí, M. Drozdzal, G. Pascual, P. Radeva, C. Malagelada, F. Azpiroz, and J. Vitrià, "Generic feature learning for wireless capsule endoscopy analysis," *Computers in biology and medicine*, vol. 79, pp. 163–172, 2016.

[27] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[28] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," 2016.

[29] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," 2016.

[30] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," *arXiv preprint arXiv:1911.05722*, 2019.

[31] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[32] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Advances in neural information processing systems*, vol. 33, pp. 22 243–22 255, 2020.

[33] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," *arXiv preprint arXiv:2103.03230*, 2021.

[34] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 271–21 284, 2020.

[35] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," 2020. [Online]. Available: https://arxiv.org/abs/2006.09882

[36] X. Chen and K. He, "Exploring simple siamese representation learning," 2020. [Online]. Available: https://arxiv.org/abs/2011.10566

[37] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *ArXiv*, vol. abs/1807.03748, 2018.

[38] M. Ye, X. Zhang, P. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6203–6212, 2019.

[39] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

[40] 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.

[41] Y. You, I. Gitman, and B. Ginsburg, "Scaling sgd batch size to 32k for imagenet training," *arXiv preprint arXiv:1708.03888*, vol. 6, no. 12, p. 6, 2017.

[42] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[43] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.