A Comparative Evaluation of Self-Supervised Methods Applied to Rock Images Classification

Van Thao Nguyen, Dominique Fourer, Desiré Sidibé, Jean-François Lecomte, Souhail Youssef

A Comparative Evaluation of Self-Supervised Methods Applied to Rock Images Classification

Van Thao Nguyen[1,2], Dominique Fourer[2] and Désiré Sidibé[2] and Jean-François Lecomte[1] and Souhail Youssef[1]

[1]*IFPEN, Ruel-Malmaison, France*

[2]*IBISC - Univ. Évry Paris-Saclay, Évry-Courcouronnes, France*

van-thao.nguyen@ifpen.fr, dominique.fourer@univ-evry.fr

Keywords: Self-Supervised Learning, Representation Learning, Digital Rock Physics Digital Rock Physics (DRP), Image Classification.

Abstract: Digital Rock Physics DRP is a discipline that employs advanced computational techniques to analyze and simulate rock properties at the pore-scale level. Recently, Self-Supervised Learning (SSL) has shown promising outcomes in various application domains, but its potential in DRP applications remains largely unexplored. In this study, we propose to assess several self-supervised representation learning methods designed for automatic rock category recognition. Hence, we demonstrate how different SSL approaches can be specifically adapted for DRP, and comparatively evaluated on a new dataset. Our objective is to leverage unlabeled micro-CT (Computed Tomography) image data to train models that capture intricate rock features and obtain representations that enhance the accuracy of classical machine-learning-based rock images classification. Experimental results on a newly proposed rock images dataset indicate that a model initialized using SSL pretraining outperforms its non-self-supervised learning counterpart. Particularly, we find that MoCo-v2 pretraining provides the most benefit with limited labeled training data compared to other models, including supervised model.

1 INTRODUCTION

Understanding fluid flow and mass transport in porous media holds paramount significance in various geological and engineering applications such as groundwater management, soil mechanics and geotechnical engineering (geological carbon storage, subsurface contaminant transport or CO_2-sequestration to name a few). However, characterizing complex rocks remains a challenging endeavor due to inherent heterogeneities observed at all scales of observation and measurement (Andrä et al., 2013). In recent decades, the advancement of synchrotron X-ray tomography has substantially enhanced the acquisition of highly precise 3D images, thereby revolutionizing the field of material research. The information acquired from the images aids in determining petrophysical and flow properties of porous media, crucial for understanding for understanding their widespread presence in soil, rock formations, and composite materials (Blunt et al., 2013). For the successful development of digital rock models, it is crucial to provide a comprehensive description and characterization of porous media such as rocks for accurate classification. Traditional approaches to rock image classification predominantly rely on manual operations, leading to both high costs and variable degrees of accuracy.

Nowadays, machine learning techniques achieve outstanding performances in various application areas, and the field of automatic classification of rock images is not an exception. Many works have been done in DRP such as using shallow neural networks to classify rock images (Guojian et al., 2013), identifying the rock granularity by a Convolutional Neural Networks (Cheng and Guo, 2017), or building a neural network to identify the rock mineral (Liu et al., 2021). Despite the success in those different applications, most of state-of-the-art methods learn features in a supervised way, and are restricted to a given specific task. Furthermore, the process of labeling rock images through micro-CT is a difficult problem which requires significant exertion, computation, and annotations from geological specialists (Karimpouli and Tahmasebi, 2019) (Shim et al., 2023). To address this challenge, one can rely on SSL methods which have shown successes in building large-scale deep-learning-based approaches to learn the underlying representations from unlabeled data (Dosovitskiy

et al., 2014) (Misra and Maaten, 2020).

SSL has recently emerged as a leading approach in achieving state-of-the-art performance in visual representation learning, eliminating the need for extensive dataset annotation. The SSL method follows a two-step process (as illustrated in Figure 1):Initially, the ConvNet is pretrained using unlabeled data, addressing what is referred to as the pretext task. Subsequently, fine-tuning on the target task using labeled data is performed in the downstream task. Notably, pretext task training is conducted autonomously, without requiring external human supervision. Consequently, this approach leverages unlabeled data to boost system performance. SSL techniques have demonstrated strong performances, particularly in the field of natural image classification, such as the ImageNet challenge (Jaiswal et al., 2020). Nonetheless, these methods are typically designed for color images, aiming to acquire color image representation. As a result, they may not be well-suited for a DRP dataset, which is often based on a grayscale conveying a specific physical information. In such datasets, images exhibit similar spatial structures, and obtaining annotated data is often more challenging compared to natural images.

In this paper, we evaluate the effectiveness of four distinct and promising SSL approaches, adapted for the automatic classification of micro-CT rock images. We have chosen four methods, namely SimCLR (Chen et al., 2020a), MoCo-v2 (Chen et al., 2020b), BYOL (Grill et al., 2020) and NNCLR (Dwibedi et al., 2021), based on their promising results in image classification tasks and their demonstrated computational efficiency, as supported by existing literature. Hence, each of these investigated models presents a unique set of pros and cons that can complement one another. The objective of this work is to provide practical insights through a comparative evaluation of these methods when adapted to the context of DRP.

The main contributions of this paper are summarized as follows:

1. We show that our proposed Self-Supervised Learning methods can yield superior representations and initialization when compared to those obtained without SSL pretraining for the task of micro-CT rock images analysis.
2. We conduct supervised fine-tuning experiments using varying fractions of labeled data. Our findings show that our pretrained representations are of greater quality than those pretrained on ImageNet. In particular, the SSL methods achieve superior classification performances when dealing with very limited labeled data.

The rest of the paper is organized as follows. Section 2 briefly describes the investigated representation learning methods, as well as the training procedures. The experimental setups and results are presented and discussed in Section 3 and the paper ends with concluding remarks in Section4.

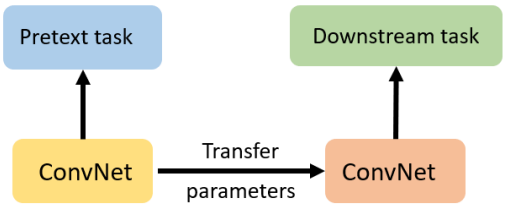


Figure 1: Illustration of the general pipeline for Self-Supervised Learning (SSL).

2 METHODS

In order to learn a robust representation of rock images, our primary objective is to ensure that the learned embeddings has the ability to capture distinct features and acquire morphological features from high-resolution X-ray micro computed tomography images of porous media. These morphological features serve as invaluable building blocks for our efforts, providing a wealth of information critical for next downstream tasks. To achieve this goal, we employ four promising self-supervised learning frameworks, namely SimCLR, MoCo-v2, BYOL, and NNCLR, specifically designed for the context of DRP

Each model comes with its own unique set of benefits such as SimCLR is renowned for its simplicity of implementation and independence from specialized architectures or memory banks (Chen et al., 2020a). SimCLR exhibits exceptional performance, particularly when used with large batch sizes,but at a higher computational cost. To address this, we also consider for Moco-v2, which uses a batch size of 256 and has demonstrated performance comparable to SimCLR on ImageNet, despite SimCLR's larger batch size of 8192 (Chen et al., 2020b).

BYOL achieves impressive performance and demonstrates robustness to the choice of image augmentations compared to contrastive methods, primarily by wiping out the reliance on negative pairs (Grill et al., 2020). Moreover, given the unique nature of our rock images dataset, the selection of appropriate transformations in SSL models is crucial. Hence, NNCLR is selected as it reduces the dependence on complex augmentations by leveraging nearest neighbors, resulting in superior performance compared to other frameworks (Dwibedi et al., 2021).

An illustration of SimCLR, MoCo-v2, BYOl and NNCLR frameworks is presented in Figure 2. Our code, inspired by (Da Costa et al., 2022), is available at `https://github.com/nguyenva04/drp_ssl`

All four frameworks consist of three primary components: (1) a data augmentation module that transforms a data sample x into two different views (x_i, x_j), which is a positive pair (different augmented views of the same image). (2) a neural encoder module which consists of 2 encoders $f_\theta(.)$ and $f_\xi(.)$ encoding the input features into a fixed dimensional embedding. (3) a projection head $g(.)$ that maps embeddings (h_i, h_j) to projections (z_i, z_j) in a latent space related to the chosen loss function. In the following subsections, we further describe each of the selected SSL methods.

2.1 SimCLR

SimCLR, the Simple Framework for Contrastive Learning of visual Representations, is a contrastive learning approach that generates augmentation-invariant embedding for input images (Chen et al., 2020a). SimCLR relies on two fundamental concepts: employing extensive data augmentation techniques to create correlated views of the same input and using a large batch size with numerous negative examples. In a batch, besides the original image and its augmented versions, all other images are considered as negative samples.

SimCLR leverages a symmetric dual-encoder architecture,wherein both encoders share parameters with each other, as illustrated in Figure (2a). During training, both encoders are updated end-to-end through backpropagation by an optimizer. The SimCLR framework uses the InfoNCE contrastive loss, also known as the normalized temperature-scaled cross-entropy loss (NT-Xent) expressed as:

$$\mathcal{L}_{i,j}(z) = -\log\left(\frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{\forall k, k\neq i}\exp(\text{sim}(z_i, z_k)/\tau)}\right), \quad (1)$$

where $\text{sim}(x, y)$ is the cosine similarity between vectors x and y, and τ is the temperature hyperparameter.

2.2 MoCo-v2

Similar to MoCo-v1 (He et al., 2020), MoCo-v2 (Chen et al., 2020b) extends the concept of instance discrimination and is designed to learn representations using a contrastive learning criterion.

For each training sample, (x_i, x_j) pairs are generated through a set of data augmentations. These pairs, x_i and x_j, are separately encoded to produce embedding z_i and z_j by both an encoder and a momentum encoder. The momentum encoder shares the same architecture as the primary encoder, as illustrated in Figure 2b. However, unlike the encoder, the momentum encoder is not backpropagated after training each mini-batch. Instead, it is updated using the parameters of the encoder, following the approach outlined in (Oord et al., 2018):

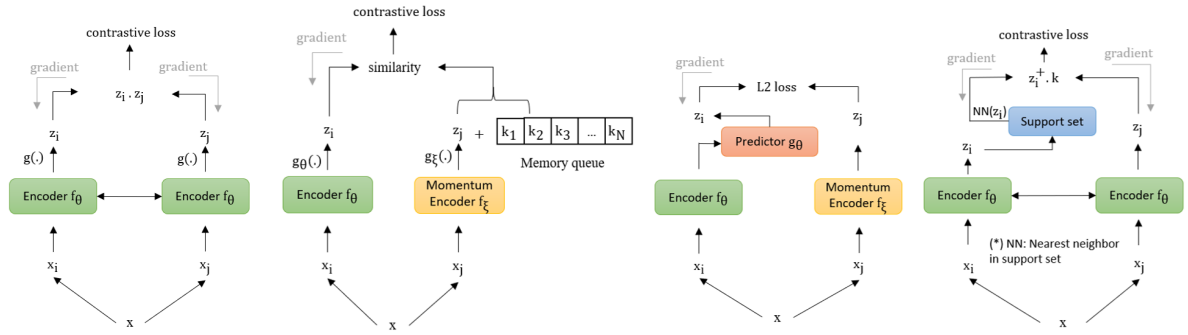
$$\xi \leftarrow m\xi + (1-m)\theta, \quad (2)$$

as represented by Eq. (2). ξ and θ are the parameters of each encoder and $m \in [0, 1]$ is the momentum coefficient. This setup allows the momentum encoder to update slowly and smoothly compared to the primary encoder. The core idea driving MoCo is to maintain a dynamic dictionary as a queue of data samples. The dictionary can be much larger than the mini-batch and easy to adjust. At the end of each training step, we update it by taking the embeddings of the momentum encoder from the current training step and concatenating them at the end of the queue. Subsequently, we discard the oldest embeddings from the queue. This approach helps maintain the consistency of the dictionary, as the oldest embeddings are often outdated and inconsistent with the new entries. MoCo-v2 demonstrates a simple improvement compared to MoCo-v1 on the ImageNet dataset by using a projection head.

2.3 BYOL

BYOL uses two neural networks for learning, namely the online network and the target network, as illustrated in Figure 2c. The online network, defined by a set of weights θ, comprises three stages: an encoder f_θ, a projector g_θ, and a predictor q_θ, all of which are trainable. On the other hand, the target network, sharing the same architecture as the online network except for the prediction head, is non-trainable and defined by a set of weights ξ. The regression targets for training the online network are provided by the target network, whose parameters ξ are updated as a moving average of the online network's parameters θ.

The mean squared error (MSE) loss is used by comparing the normalized prediction $\overline{q_\theta}(z_i)$ generated by the online network and the projection $\overline{z_\xi}$ produced by the target network. In contrast to approaches like MoCo and SimCLR, which use negative samples to prevent collapse, BYOL takes a different approach. BYOL encourages the online projection to encode more and more information by adding a predictor to the online network and using the moving average of the online network parameters as the target network. This strategy helps avoid collapsed solutions, such as constant representations.



(a) SimCLR (b) MoCo-v2 (c) BYOL (d) NNCLR

Figure 2: Different architecture pipelines for self-supervised learning.

2.4 NNCLR

In contrast to other mentioned self-supervised learning (SSL) methods, NNCLR (Dwibedi et al., 2021) introduced a novel concept regarding the selection of positive samples. Instead of generating positive samples for an image through augmentation, NNCLR demonstrated that positive samples can be obtained by using the nearest neighbors of the sample in the support set of embeddings, creating more diversified positive pairs. NNCLR acquired positive samples by determining the nearest neighbors in terms of Euclidean distance of a given sample within the learned latent space of the dataset. This approach offers more semantic class-wise variations compared to pre-defined transformations, which tend to provide more geometric information.

NNCLR shares similarities with SimCLR in terms of its straightforward architecture and with MoCo for the utilization of dynamic dictionary as a support set (first-in-first-out), as shown in Figure 2d. This support set is dynamically updated at the end of each training step by concatenating the current embeddings at the end of the queue. Unlike approaches that require extensive computational resources due to a diverse set of negative samples, NNCLR does not heavily rely on predefined data augmentation. It has the capability to establish connections among multiple samples that may potentially belong to the same semantic class.

3 EXPERIMENTAL RESULTS

In this section, we conduct several experiments using our own DRP Dataset. We begin by introducing the dataset, followed by pre-training adaptation, pre-training protocal, and transfer learning strategies. Then, we compare the SSL models with a supervised baseline method (ResNet-50) (Deng et al., 2009).

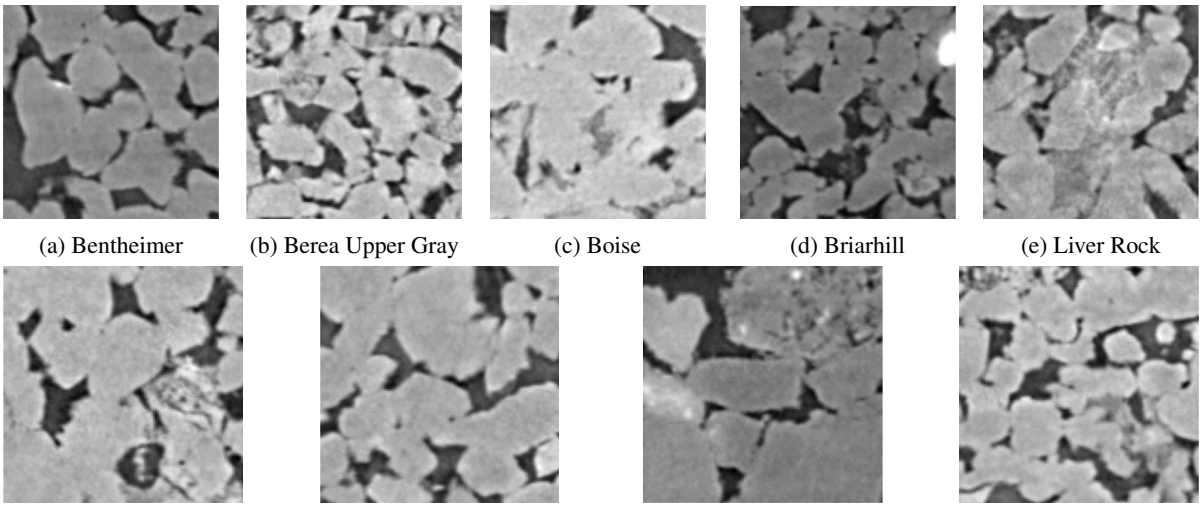
3.1 Dataset

In this paper, we use a subset of a larger collection of high-resolution 3D images of rock samples obtained through a campaign of IFP Energies Nouvelles (IFPEN). The images were acquired using tomography technology, which produces a series of 3D volumes showing the spatial distribution of the pore space in a rock sample. The dataset comprises large tomograms of fifty sandstone samples, each measuring 1100 x 1100 x 2800, are divided into smaller Regions of Interest (ROI) to create a dataset. These large tomograms are divided into ROI, each consisting of 128×128 pixels, capturing various structures present in the larger tomograms. The creation of each ROI involves randomly cutting it in any of the three dimensions. As illustrated in Figure 3, there are nine distinct sandstones in the dataset, and the number of images varies according to the sandstone. Our dataset consists of 62.500 images, with each sample producing 1.250 images. To enhance data quality, the dataset undergoes a cleaning process to remove outliers and artifacts, involving an examination of the average gray level of each image. Finally, for training and testing purposes, the dataset is split into two distinct sets in an 8:2 ratio.

The characteristic quantities for the sandstones used in this paper are listed in Table 1.

3.2 Contrastive Pretraining for DRP Adaptation

We use ResNet (Deng et al., 2009) as ConvNet for all our SSL models. This decision enables a more objective comparison of the representations learned by various approaches. Our experiments use the entire DRP training dataset, and to leverage potential con-



(a) Bentheimer (b) Berea Upper Gray (c) Boise (d) Briarhill (e) Liver Rock

(f) Castlegate (g) Idaho Gray (h) Leopard Sanstone (i) Michigan

Figure 3: A representative example from each class in the DRP dataset.

Table 1: Characteristics quantities of sandstone samples.

Type	$Porosity_{exp}$	$Permeability_{exp}$ (mD)	Nb samples
Bentheimer	24%	2000	6
Berea Upper Gray	21%	390	6
Boise	28%	1800	5
Briarhill	24%	3500	5
Castlegate	28%	1000	6
Idaho Gray	30%	6000	6
Leopard Sandstone	21%	1200	4
Michigan	21%	900	6
Liver Rock	24%	1050	6

vergence advantages, we initialize the models with ImageNet weights. The use of ImageNet weights is advantageous due to their widespread availability, eliminating the need for additional computational costs during model initialization. Given that our data is in grayscale, a specific initialization step is implemented. We initialize the weights of the first hidden layer by computing the mean of the pre-trained weights from the RGB channels (Ahmad and Shin, 2021).

We adapt our data augmentation strategy to create views suitable for the rock classification task. Traditional data augmentations commonly used in SSL for natural images are unsuitable for our dataset. For example, color jittering and random grayscale transformations are ineffective for micro-CT grayscale images, as these techniques do not yield significant improvements within the context of the DRP dataset. Instead, our augmentation strategy includes horizontal flipping, vertical flipping, Gaussian blur, and coarse dropout, as shown in Figure 4. These transformations are specifically chosen to enhance the representation learning process for the DRP dataset.

3.3 Pretraining Protocol

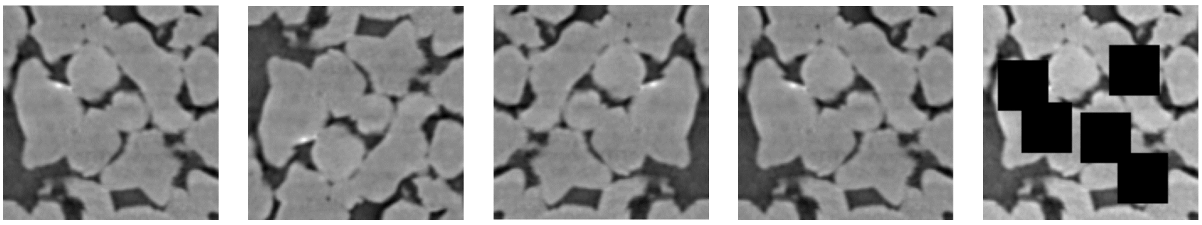
For SimCLR, we used the LARS optimizer to stabilize the pre-training, setting the learning rate of 0.01, a temperature parameter $\tau = 0.2$, and a batch size of 128. A linear warmup is applied for the first 10 epochs, followed by learning rate decay using the cosine decay schedule without restart.

For MoCo-v2, BYOL, NNCLR, and the baseline model, we used the SGD optimizer with a weight decay of 10^{-5} and a momentum value of 0.9. MoCo used a fixed momentum update coefficient of 0.996, while BYOL used an exponential moving average of momentum, starting at 0.996 and reaching 1.

The batch size is set to 128 for MoCo-v2 and 64 for BYOL, NNCLR, and the baseline. The training of SimCLR, BYOL, and NNCLR is conducted on two NVIDIA Tesla V100-PCIE-16GB GPUs, while MoCo-v2 is trained on a single NVIDIA RTX A2000 8GB GPU. MoCo-v2 and NNCLR use a dynamic dictionary of size 4096. Further details, including the number of epochs and training times for our models, are presented in Table 2.

Table 2: Training time.

Framework	nb. epochs	times
SimClr	100	4h50
MoCo-v2	50	4h30
BYOL	50	9h
NNCLR	50	6h



(a) Original Image (b) Vertical Flip (c) Horizontal Flip (d) Gaussian Blur (e) Cutout

Figure 4: Illustration of different data transformations.

3.4 Transfer Learning

In this section, we evaluate the quality of representations by fine-tuning on downstream tasks. The fine-tuning process involves using different fractions of labeled training data, where the label fraction denotes the proportion of retained data labels during training. For example, a model trained with a 1% label fraction has access to only 1% of all labels, while the remaining 99% are hidden. This simulation is reflective of real-world scenarios where large amounts of data remain unlabeled, and only a small fraction of well-labeled data is available for supervised training. In our experiments, we use label fractions of 1%, 10%, and 50% for the training set in our dataset.

In order to evaluate the quality of the representations, we use two distinct approaches. Firstly, we freeze the backbone model's parameters and conduct training of a linear classifier using the labeled data. Secondly, we unfreeze all layers and carried out an End-to-End fine-tuning of the entire model, again using the labeled data. We also conduct a baseline for comparison, contrasting classical supervised methods with self-supervised models. This baseline uses ImageNet-pretrained models without undergoing contrastive pretraining but shares the same backbone network as the contrastive models.

The linear evaluation, End-to-End, and baseline approaches all share identical configurations. Our model undergoes training for 50 epochs, utilizing a batch size of 64, an initial learning rate of 0.01, and a scheduled learning rate decrease of 0.1 after every 20 epochs. The SGD optimizer is employed for optimization.

3.5 Results

Through the comparison of classifiers relying on pretrained networks such as SimCLR, MoCo-v2, BYOL and NNCLR, as opposed to a classifier initialized with ImageNet, we aim to show how well a pretrained neural network provides useful representations. We started with linear evaluation (Oord et al., 2018) and we visualize the performance of different approaches by using linear evaluation in Table 3 and by End-to-End in Table 4 at various fractions.

The results in Table 3 show that MoCo-v2 outperforms the other self-supervised learning models and the supervised baseline on all metrics (recall, precision, F-score, and top-1 accuracy) for all label fractions (1%, 10%, and 50%). Specifically, MoCo-v2 achieves a top-1 accuracy of 63.64% with a label fraction of 1%, 69.09% with a label fraction of 10%, and 72.96% with a label fraction of 50%. This is significantly better than the supervised baseline, which achieves a top-1 accuracy of 43.29% with a label fraction of 1%, 53.56% with a label fraction of 10%, and 55.75% with a label fraction of 50%. The other self-supervised learning models (SimCLR, BYOL and NNCLR) also outperform the supervised baseline, but they do not perform as well as MoCo-v2. These results are promising since it shows the ability of SSL models to learn representations that capture rich data semantics and informative features.

Table 3: ResNet-50 Linear Evaluation.

Method name	Recall	Precision	F-Score	top-1 Acc
SimClr 1%	58.48	57.65	58.06	58.48
MoCo-v2 1%	61.73	62.35	62.04	**63.64**
BYOL 1%	56.07	55.04	55.55	56.07
NNCLR 1%	55.17	55.58	55.37	55.17
Baseline 1%	43.29	39.48	41.30	43.29
SimClr 10%	64.60	63.90	64.25	64.60
MoCo-v2 10%	67.64	67.20	67.42	**69.09**
BYOL 10%	64.54	64.32	64.45	64.54
NNCLR 10%	66.74	66.56	66.65	66.74
Baseline 10%	53.56	51.99	52.76	53.56
SimClr 50%	69.18	68.84	69.01	69.18
MoCo-v2 50%	71.91	71.56	71.73	**72.96**
BYOL 50%	69.05	68.66	68.85	69.05
NNCLR 10%	70.34	69.79	70.06	70.34
Baseline 50%	55.75	55.31	55.53	55.75

Subsequently, an empirical investigation is conducted to assess the usefulness of our SSL pretrained models in providing enhanced representations through End-to-End strategies. Our findings indicate that pre-training using all SSL fine-tuning models significantly improve the accuracy with label-efficiency for micro-CT images of rock classifica-

Table 4: ResNet-50 End-to-End.

Method name	Recall	Precision	F-Score	top-1 Acc
SimClr 1%	60.78	61.10	60.94	60.78
MoCo-v2 1%	66.53	68.28	67.39	**66.53**
BYOL 1%	62.05	64.40	63.20	62.05
NNCLR 1%	60.40	62.05	61.21	60.40
Baseline 1%	58.21	58.36	58.29	58.21
SimClr 10%	75.58	75.30	75.44	75.58
MoCo-v2 10%	79.84	79.69	79.77	**79.84**
BYOL 10%	77.84	77.80	76.81	77.84
NNCLR 10%	75.43	75.43	75.43	75.43
Baseline 10%	73.81	73.73	73.77	73.81
SimClr 50%	89.42	89.43	89.43	89.42
MoCo-v2 50%	90.05	90.20	90.13	90.37
BYOL 50%	90.16	90.11	90.14	90.16
NNCLR 50%	90.45	90.46	90.46	90.45
Baseline 50%	90.55	90.55	90.55	**90.55**

Table 5: Accuracy improvements achieved by SSL pretrained models against model without SSL pretrained on the DRP dataset.

SSL-framework	SSL-pretrained	ImageNet-pretrained	1%	10%	50%
SimClr	Linear Model	Linear Model	0.3508	0.2061	0.2409
MoCo-v2	Linear Model	Linear Model	0.4700	0.2900	0.3087
BYOL	Linear Model	Linear Model	0.2952	0.2050	0.2386
NNCLR	Linear Model	Linear Model	0.2744	0.2461	0.2617
SimClr	End-to-End	End-to-End	0.042	0.024	-0.009
MoCo-v2	End-to-End	End-to-End	0.1429	0.082	-0.002
BYOL	End-to-End	End-to-End	0.0660	0.055	-0.004
NNCLR	End-to-End	End-to-End	0.0376	0.022	-0.001

tion, self-supervised models outperform the supervised baseline by a larger margin when trained on smaller datasets. In the results of fine-tuning, as shown in Table 5, MoCo-v2 outperforms again the supervised baseline and the other SSL models for small label fractions (1%, 10%). The other SSL models are similar to MoCo-v2 by outperforming the baseline model when trained on small datasets. These results suggest that self-supervised model yields proportionally larger gains when End-to-End with fewer label samples and less significant at larger label fraction. This result is consistent with (Chen et al., 2020a), which shows that self-supervised models trained on smaller fractions of the ImageNet dataset gain better improvement.

Our test results confirm that MoCo-v2 is a high-performing SSL model. We also observe that BYOL performs worse than expected, despite its high accuracy on ImageNet. One possible explanation is that MoCo-v2 may learn better representations with our chosen set of strong data augmentations. (Huang et al., 2022) has shown that MoCo-v2 catches up to BYOL in terms of linear accuracy when trained with more complex data augmentations. However, it is important to note that we trained BYOL with a smaller batch size (64), while BYOL has been shown to achieve optimal results with a larger batch size of 4096, reaching 74.3% accuracy after training on 512 TPUs. Therefore, it is not fair to directly compare the performance of MoCo-v2 and BYOL in this case.

The additional results for each class are presented in Table 6. It becomes evident that Idaho Gray stands out as the most easily discernible rock type, while Leopard proves to be the most challenging for recognition. Notably, compared to the other classes in our dataset, the Leopard rock has the least amount compared to other classes in our dataset, which could potentially explain its comparatively lower accuracy score. In limited labeled DRP data scenarios, SSL frameworks outperform supervised methods significantly. However, when abundant labeled data is available, the improvement from SSL models in the DRP context diminishes compared to situations with fewer data samples.

4 CONCLUSION

In this study, we have discovered that SSL frameworks can provide suitable representations for the analysis of micro-CT images of rock. These representations significantly enhance the performance of downstream tasks compared to traditional supervised learning methods across all label fractions. The challenges posed by the scarcity and cost associated with manual annotation in the domain of DRP make it difficult to acquire datasets of a scale comparable to those in well-established computer vision domains. Our success in showcasing performance improvements over traditional supervised learning methods, particularly in scenarios with limited labeled data, holds the potential for broader applications in micro-CT image analysis, encompassing both 2D and 3D representations characterized by intricate textures. A pretrained network tailored for DRP can be used for different purposes, including classification, segmentation, and the estimation of rock properties.

Based on our analysis, we anticipate that increased computational resources allocated to model training could result in improved performance. Additionally, we identify the potential for further research, particularly in the exploration of various downstream tasks such as regression which are of interest for 3D tomography of rock images.

REFERENCES

Ahmad, I. and Shin, S. (2021). An approach to run pre-trained deep learning models on grayscale images. In *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 177–180.

Table 6: Additional results: Accuracy per class by End-to-End.

Method name	Bentheimer	Berea UG	Boise	Briarhill	Castlegate	Idaho Gray	Leopard	Michigan	Liver Rock
SimClr 1%	71.70	66.40	**72.62**	55.61	73.63	73.61	35.67	41.35	49.10
MoCo-v2 1%	**71.83**	**73.60**	69.42	56.10	**77.44**	**85.37**	**42.59**	44.16	59.16
BYOL 1%	59.28	62.20	58.29	**61.46**	74.90	83.43	30.66	**54.44**	**62.61**
NNCLR 1%	66.10	53.40	58.85	56.10	70.50	84.57	28.86	**54.44**	59.33
Baseline 1%	65.42	59.27	58.53	55.85	73.43	79.49	35.97	36.54	45.62
SimClr 10%	84.65	81.27	76.86	70.91	89.92	85.91	47.40	72.55	60.54
MoCo-v2 10%	**86.98**	**83.73**	**81.10**	**76.12**	93.79	87.44	**56.41**	**73.88**	**70.84**
BYOL 10%	83.11	82.33	71.74	72.92	**90.12**	**97.64**	49.00	68.27	64.41
NNCLR 10%	80.51	81.73	78.78	69.79	90.05	86.17	54.61	63.73	66.15
Baseline 10%	86.45	79.80	78.70	67.31	89.65	84.70	48.90	68.67	60.87
SimClr 50%	90.99	92.80	89.43	87.42	97.13	92.92	80.66	88.31	84.48
MoCo-v2 50%	90.12	92.07	**89.60**	88.86	96.53	91.98	**83.53**	90.05	**86.09**
BYOL 50%	93.93	93.33	86.55	**89.42**	97.66	**95.73**	82.16	89.31	82.01
NNCLR 10%	93.39	90.93	88.39	87.82	**97.93**	94.46	81.66	**90.72**	85.02
Baseline 50%	**94.26**	**93.47**	88.63	87.74	97.60	95.12	78.86	88.71	84.15

Andrä, H., Combaret, N., Dvorkin, J., Glatt, E., Han, J., Kabel, M., Keehm, Y., Krzikalla, F., Lee, M., Madonna, C., et al. (2013). Digital rock physics benchmarks—part i: Imaging and segmentation. *Computers & Geosciences*, 50:25–32.

Blunt, M. J., Bijeljic, B., Dong, H., Gharbi, O., Iglauer, S., Mostaghimi, P., Paluszny, A., and Pentland, C. (2013). Pore-scale imaging and modelling. *Advances in Water resources*, 51:197–216.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020a). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Chen, X., Fan, H., Girshick, R., and He, K. (2020b). Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Cheng, G. and Guo, W. (2017). Rock images classification by using deep convolution neural network. In *Journal of Physics: Conference Series*, volume 887, page 012089. IOP Publishing.

Da Costa, V. G. T., Fini, E., Nabi, M., Sebe, N., and Ricci, E. (2022). solo-learn: A library of self-supervised methods for visual representation learning. *Journal of Machine Learning Research*, 23(56):1–6.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Dosovitskiy, A., Springenberg, J. T., Riedmiller, M., and Brox, T. (2014). Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27.

Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., and Zisserman, A. (2021). With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284.

Guojian, C., Wei, M., Xinshan, W., Chunlong, R., and Junxiang, N. (2013). Research of rock texture identification based on image processing and neural network. *J Xi'an Shiyou Univ (Nat Sci Edition)*.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Huang, J., Kong, X., and Zhang, X. (2022). Revisiting the critical factors of augmentation-invariant representation learning. In *European Conference on Computer Vision*, pages 42–58. Springer.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. (2020). A survey on contrastive self-supervised learning. *Technologies*, 9(1):2.

Karimpouli, S. and Tahmasebi, P. (2019). Segmentation of digital rock images using deep convolutional autoencoder networks. *Computers & geosciences*, 126:142–150.

Liu, Y., Zhang, Z., Liu, X., Wang, L., and Xia, X. (2021). Ore image classification based on small deep learning model: Evaluation and optimization of model depth, model structure and data size. *Minerals Engineering*, 172:107020.

Misra, I. and Maaten, L. v. d. (2020). Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Shim, M. S., Thiele, C., Vila, J., Saxena, N., and Hohl, D. (2023). Content-based image retrieval for industrial material images with deep learning and encoded physical properties. *Data-Centric Engineering*, 4:e21.