Multi-Task Continued Pre-training: A Sequential (Contrastive-Pretext) Framework for Robust Medical Image Classification

1st Manuel Gortarez
Dept. of Mathematics
University of Sonora
Hermosillo, México
0009-0000-8519-6396

2nd Alan Torres
Dept. of Mathematics
University of Sonora
Hermosillo, México
0009-0007-0173-075X

3rd Omar Pacheco
Dept. of Mathematics
University of Sonora
Hermosillo, Mexico
0009-0007-3331-0741

4th Dante Tostado-Cortes
Dept. of Mathematics
University of Sonora
Hermosillo, México
0009-0009-8953-874X

5th Fernando Bórquez
Dept. of Mathematics
University of Sonora
Hermosillo, Mexico
0009-0009-3600-0444

***Abstract*—Domain shift remains a critical challenge for robust medical image classification. In this work for the Medical Domain Adaptation (MeDA) Challenge 2025, we propose a novel sequential self-supervised framework to learn domain-invariant representations. Our method first uses contrastive learning to capture global features and then employs a multi-pretext stage to refine them with local, structural information. This synergistic approach yields a model with significantly improved stability and competitive accuracy on MedMNIST v2 datasets, particularly in data-scarce, cross-domain settings. Our results underscore the framework's effectiveness in creating more generalizable and reliable classifiers for real-world medical applications.**

***Index Terms*—Self-Supervised Learning, Robust Generalization, Medical Imaging, Cross-Domain Classification, MedMNIST v2**

I. Introduction

The use of machine learning in medical imaging has demonstrated great potential for improving diagnostic accuracy and clinical decision-making. However, one of the primary obstacles in this field is the limited availability of large, diverse, and well-annotated datasets [1]. Unlike domains such as natural image recognition, where large-scale datasets like ImageNet have driven progress, medical data collection faces several challenges, including privacy concerns, annotation costs, and the scarcity of expert-labeled samples. These limitations hinder the development of models capable of achieving strong performance and *robust generalization* across diverse clinical scenarios.

To overcome these challenges, research efforts have increasingly focused on methods that enable models to adapt to unseen domains and datasets. In particular, *domain adaptation* and *cross-domain generalization* aim to improve the transferability of learned representations, allowing models trained in one context to remain effective in others. Initiatives such as the *Medical Domain Adaptation (MeDA) Challenge 2025* encourage the exploration of algorithms capable of learning domain-invariant features for medical image classification, highlighting the importance of generalizable approaches in data-limited environments [2], [3].

Generalization, in this context, refers to a model's ability to maintain performance across datasets that differ in acquisition protocols, imaging devices, or patient populations. It is a key indicator of model robustness and reliability in real-world clinical applications. Without sufficient generalization, models that perform well on their training data often fail when applied to new hospitals or imaging modalities. Recent studies have emphasized this issue, evaluating how well self-supervised learning (SSL) models generalize beyond their training distributions [2]. Approaches such as variational encoding with linear-dependency regularization and domain-generalization frameworks without domain labels have been proposed to improve cross-domain performance, yet challenges persist due to the inherent variability of medical data.

Among the strategies to enhance model generalization, Self-Supervised Learning (SSL) has emerged as a particularly promising paradigm. Unlike supervised learning, which depends heavily on labeled data, SSL leverages unlabeled samples by formulating pretext tasks that allow models to learn meaningful visual representations autonomously. These representations can later be fine-tuned for downstream tasks, such as classification or segmentation, with minimal labeled data. In medical imaging, where expert annotations are expensive and time-consuming, SSL provides an efficient way to exploit the vast amounts of unlabeled data available. Recent work [2], [4] demonstrates that SSL not only improves representation quality but also enhances robustness and generalization across diverse imaging modalities and diagnostic tasks. By capturing structural, spatial, and contextual information from medical images, SSL-based models pave the way for scalable, reliable, and transferable solutions in medical image analysis.

II. Approach

Our approach is based on sequentially integrating two distinct paradigms of Self-Supervised Learning (SSL) that complement each other. The proposed process comprises an initial pre-training stage followed by a continued pre-training stage, strategically designed to enhance the representational capacity of the final encoder.

A. Initial Pre-Training (MoCo)

In the first stage, we train a ResNet-18 encoder using Momentum Contrast (MoCo) [5]. The objective of this phase is to establish a robust baseline feature extractor through a contrastive, instance-discrimination objective. This task encourages the model to learn representations that remain highly invariant to common augmentations (e.g., cropping, rotation, color jitter) while capturing global semantic similarity. As a result, it provides a powerful and generalized feature space serving as the foundation for the subsequent refinement phase.

B. Continued Pre-Training (Multi-Pretext)

We then subject the MoCo-initialized encoder to a secondary self-supervised training stage, employing a multi-task pretext framework inspired by [4].

The purpose of this stage is to refine and specialize the general-purpose, invariant features obtained from MoCo. While MoCo focuses on global similarity, the pretext-based framework enforces the learning of explicit contextual and structural priors intrinsic to the visual domain. To accomplish this, we attach multiple task-specific predictive heads to the shared backbone and train them stochastically on three distinct pretext tasks:

- Colorization: A convolutional decoder predicts the original color image from a grayscale input, compelling the model to capture fine-grained texture and surface information.
- Patch Prediction (Inpainting): A second decoder reconstructs a masked central patch from its surrounding context, enabling the model to infer local contextual relationships.
- Jigsaw Puzzle Solving: A classification head predicts the correct permutation of a shuffled 2×2 grid of image patches, enforcing reasoning over spatial relationships and global structure.

During this refinement stage, the loss from a randomly selected pretext task is backpropagated to update the parameters of both the task-specific head and the shared MoCo backbone.

C. Final Representation

As we established before, the motivation behind this sequential design is to synergistically combine two complementary SSL paradigms:

The MoCo encoder yields a global, invariant representation, excelling at encoding what an image is, regardless of its presentation.

The Multi-Pretext encoder yields a local, structural representation, teaching the model how the components of an image are spatially and contextually organized.

The sequential integration of these methods represents a deliberate effort to construct a hierarchical feature extractor. The final encoder is not merely the outcome of a single SSL method but a hybrid representation learner. This method embodies both the powerful global invariance of contrastive learning and the high-fidelity contextual awareness derived from pretext-based learning. Consequently, this enriched feature space allows for a more efficient learning of new tasks and generalization across diverse datasets.

III. Results

This section details the empirical results of our experiments, comparing the generalization performance of our Self-Supervised Learning (SSL) model against standard ImageNet-pretrained baselines in a few-shot learning context.

A. Setup

To evaluate the quality of the learned feature representations, we employed a linear probing protocol. We compared three distinct backbone architectures:

- MTCP: The proposed framework, with a ResNet-18 backbone.
- ImageNet-R18: A standard ResNet-18 backbone pretrained on the ImageNet dataset.
- ImageNet-R50: A standard ResNet-50 backbone pretrained on the ImageNet dataset.

For all experiments, the downstream task was classifying 3 distinct anatomical landmarks of the GI tract, using the Kvasir dataset [6]. The backbones of all models were frozen, and only a new, randomly-initialized linear classification head was trained.

We simulated a few-shot scenario by training this linear head on a small support set sampled from the training data, with $K = \{1, 5, 10, 20, 50\}$ images per class. Model performance was then evaluated on a fixed, held-out test set. The reported results are the mean accuracy and standard deviation across multiple runs with different random seeds to ensure statistical robustness.

B. Few-Shot Learning Performance

The aggregated performance of the three models across all K-shot values is presented in Table I. Our empirical results reveal a nuanced but critical relationship between pretraining strategy, data availability, and generalization robustness. The findings validate our hypothesis that a hybrid SSL approach yields a more stable and dependable feature representation in data-scarce, cross-domain scenarios.

Performance vs. Data Availability: As expected, the test accuracy for all three models demonstrates a positive correlation with the number of shots (K) in the support set. This confirms that the linear classifier can leverage additional examples to build a more robust decision boundary, regardless of the pretraining method.

Superior Stability in One Shot: In the most data-constrained setting (1-shot), a key finding emerges not from the mean accuracy, but from the variance across different data splits (seeds).

The ImageNet-pretrained models (R18 and R50) exhibit high instability, with large standard deviations of 7.77 and 12.53, respectively. This suggests their performance is highly sensitive to the specific support image selected, making them unreliable for 1-shot generalization.

TABLE I
MEAN TEST ACCURACY AND STANDARD DEVIATION FOR THE FEW-SHOT LINEAR PROBING TASK ON THE KVASIR DATASET.

K (Shots)	MTCP Mean Acc. (±Std)	ImageNet-R18 Mean Acc. (±Std)	ImageNet-R50 Mean Acc. (±Std)
1	56.08% (±4.19)	65.12% (±7.77)	63.06% (±12.53)
5	76.68% (±5.88)	82.59% (±3.84)	85.18% (±2.54)
10	84.92% (±1.95)	88.77% (±4.16)	91.10% (±1.94)
20	85.38% (±2.68)	92.69% (±2.09)	93.55% (±2.88)
50	86.64% (±4.45)	93.09% (±3.29)	94.55% (±1.26)

In sharp contrast, our MTCP model yields the most stable and consistent results, with a significantly lower standard deviation of only 4.19.

We attribute this superior stability directly to our sequential pretraining strategy. The features from ImageNet, while powerful, are optimized for natural images. In a 1-shot cross-domain scenario, these features can be "brittle," over-fitting to the specific artifacts of a single medical image. Our model, however, combines the global invariance of MoCo with the contextual and structural priors from pretext tasks. This hybrid representation is less dependent on superficial sample features and more attuned to the underlying structure of the medical images, providing a robust generalization baseline even when labeled data is minimal.

Competitive Performance and Robustness (K=5 to K=50): As the number of shots increases, the MTCP model proves to be highly competitive and, in some cases, superior to its direct architectural counterpart, the ImageNet-R18.

While the ImageNet baselines eventually lead in mean accuracy at K=50, our detailed seed analysis (present in the supplementary material/appendix) shows that the proposed model outperforms the ImageNet-R18 baseline in specific data splits. For instance, in Table II at K=5 (SEED=43), our SSL model achieved 80.73% accuracy, surpassing the ImageNet-R18's 76.74%.

Crucially, at K=10, MTCP not only achieves a competitive mean accuracy (84.92%) but also demonstrates vastly superior stability (Std. Dev. 1.95) compared to the more erratic ImageNet-R18 (Std. Dev. 4.16).

TABLE II
DETAILED PERFORMANCE BREAKDOWN FOR SEED=43, ILLUSTRATING A SPECIFIC RUN WHERE MTCP OUTPERFORMS IMAGENET-R18 AT K=5.

K (Shots)	MTCP	ResNet-18	ResNet-50
1	59.14%	53.82%	67.44%
5	80.73%	76.74%	81.73%
10	87.71%	91.69%	93.69%
20	89.70%	94.68%	97.01%
50	93.69%	95.02%	96.01%

These findings suggest a powerful conclusion: while supervised pretraining on ImageNet remains a strong baseline, its generalization in data-scarce medical scenarios is inconsistent. Our SSL approach, by learning both global and local representations, provides a more dependable and robust foundation. The high variance of the ImageNet models, contrasted with the consistency of our SSL model, strongly supports our paper's central thesis: this hybrid SSL framework is a more reliable method for building classifiers in real-world medical environments where data scarcity and domain shift are the primary challenges.

ACKNOWLEDGMENT

We would like to express our gratitude to the **ACARUS team** at the **University of Sonora** for providing us computational resources, by letting us access the YUCA cluster, as well as technical support. Their assistance was essential for the development and execution of this project.

REFERENCES

[1] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren, "Preparing medical imaging data for machine learning," *Radiology*, vol. 295, pp. 4–15, Feb. 2020.

[2] V. Bundele, K. Sarıtaş, B. Kargi, *et al.*, "Evaluating self-supervised learning in medical imaging: A benchmark for robustness, generalizability, and multi-domain impact," *arXiv preprint arXiv:2412.19124*, Jul 2025.

[3] J. Yang, R. Shi, D. Wei, *et al.*, "Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification," *Scientific Data*, vol. 10, no. 1, p. 41, 2023.

[4] S. A. Nezhad, G. Tajeddin, T. Khatibi, and M. Sohrabi, "Self-supervised learning framework for efficient classification of endoscopic images using pretext tasks," *PLOS One*, vol. 20, no. 5, p. e0322028, 2025.

[5] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," 2020.

[6] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, MMSys'17, (New York, NY, USA), pp. 164–169, ACM, 2017.