Finetuning Pretrained Models for Compressed Dermatology Image Analysis

Sonnet Xu
Stanford University
Stanford, CA
sonnet@stanford.edu

Eric Cui
Stanford University
Stanford, CA
ericcui@stanford.edu

Aaron Fanous
Stanford University
Stanford, CA
aron7628@stanford.edu

Vicky Bikia
Stanford University
Stanford, CA
bikia@stanford.edu

Abstract

High-resolution dermoscopic images are essential for accurate skin-lesion diagnosis, yet they are costly to store and process—constraints that are especially acute in teledermatology, where bandwidth and hardware are limited. This study investigates how vision transformers (ViT) and self-supervised models behave when finetuned on compressed, downsized dermoscopic images. Using the ISIC 2019 dataset, we evaluate ViT, SimCLR, and DINOv2 at multiple input resolutions while tracking computational demands. We find that reducing resolution markedly lowers memory and FLOPs with only a minor loss in classification accuracy. The ImageNet-pretrained ViT remains the top performer across resolutions, and DINOv2 attains comparable accuracy after modest hyperparameter tuning, underscoring both architectures as strong candidates for low-resource dermatology applications. These results provide practical guidance for deploying efficient AI in bandwidth-constrained clinical settings and offer insights transferable to other medical-imaging tasks.

1. Introduction

Dermoscopic imaging has emerged as a fundamental tool in dermatological diagnostics, particularly for the early detection of melanoma and other skin cancers. These images, often captured at high resolutions, enable clinicians and AI models alike to assess subtle textural and color-based patterns indicative of disease. However, the computational and storage demands of high-resolution imaging pose a significant barrier to scalable diagnostic solutions, especially in resource-constrained environments such as mobile health clinics or teledermatology platforms. In these settings, image compression becomes a practical necessity.

This project proposes a strategy to mitigate the trade-off between image resolution and diagnostic performance by leveraging pretrained foundation models (FMs). Specifically, we investigate the diagnostic efficacy and computational efficiency of finetuning FMs on downsized dermoscopic images from the ISIC 2019 dataset. We aim to answer whether reduced-resolution inputs—such as 56×56 or 112×112—can be used without significantly sacrificing model performance in tasks like skin lesion classification.

This study addresses the following core question: How does finetuning pretrained foundation models on downsized dermoscopic images impact diagnostic performance in dermatology tasks such as skin cancer classification?

To answer this question, our study focuses on four primary objectives:

1. Finetune pretrained foundation models on ISIC 2019 images resized to varying resolutions (e.g., 56×56, 112×112, 224×224).

2. Evaluate model performance on downstream dermatology tasks, including lesion classification.

3. Measure computational efficiency, specifically floating point operations (FLOPs) and GPU memory usage, for models operating on compressed images.

4. Conduct a targeted, lightweight hyperparameter sweep to assess how much additional performance can be unlocked with minimal tuning.

2. Related Work

2.1. Foundation Models in Medical Imaging

The application of large pretrained models, including vision transformers, to medical imaging tasks has shown promise in recent years. Vision Transformer (ViT) models, originally introduced by Dosovitskiy et al. [9], have demonstrated strong performance in general vision benchmarks and have been adapted for specialized domains such as dermatology [3] and pathology [6]. These studies highlight the potential of ViTs to capture intricate patterns in medical images, such as those found in dermoscopic datasets like ISIC 2019. However, the transferability of these models to medical domains often requires careful finetuning due to the domain shift between general-purpose image datasets and highly specialized medical imaging data [2]. For instance, [2] explored the use of large-scale pretrained models in medical imaging, emphasizing the need for domain-specific adaptation to maintain diagnostic accuracy in tasks like skin lesion classification.

Recent advancements in foundation models have also focused on improving their robustness to domain-specific challenges, such as varying imaging conditions in teledermatology. [15] investigated the transfer learning capabilities of pretrained models in medical imaging, finding that finetuning on domain-specific datasets, such as ISIC 2019, significantly improves performance over direct application of pretrained weights. This is particularly relevant in resource-constrained environments, where computational efficiency and model performance must be balanced.

2.2. Self-Supervised Representation Learning

Self-supervised learning (SSL) has emerged as a powerful paradigm for learning robust representations without extensive labeled data, which is particularly valuable in medical imaging where annotated datasets are often scarce. Methods like SimCLR [5] utilize contrastive learning to create generalizable feature representations by maximizing agreement between augmented views of the same image. DINO [4] similarly employs a self-distillation approach, achieving competitive performance in low-label scenarios. DINOv2, a more recent iteration, incorporates improvements in training stability and scale [14], making it a candidate for foundation modeling in complex datasets like ISIC 2019.

Recent work has explored the use of SSL in medical imaging, particularly for improving generalization across institutions and imaging modalities [2]. For example, [1] demonstrated that SSL-pretrained models can achieve strong performance in dermatology tasks by learning transferable features that are robust to variations in image acquisition. This suggests SSL approaches may be particularly relevant for teledermatology, where models must generalize across diverse imaging conditions, such as those encountered in mobile health clinics, which aligns with our study's focus on evaluating SSL models on downsized dermoscopic images to enable efficient AI deployment in low-resource clinical settings.

2.3. Resolution and Computational Efficiency

Resolution plays a key role in medical imaging, where fine-grained visual features may be essential for accurate diagnosis. Prior studies have investigated the trade-offs between input resolution, performance, and computational cost [16, 18]. In the context of foundation models, particularly ViTs, the quadratic scaling of memory with image size presents unique challenges.

In dermatology, where fine-grained visual features are critical for accurate diagnosis, the impact of resolution reduction is less straightforward. Our study is motivated by the largely unexplored application of such findings to ViTs and SSL models. We focus on finetuning ViTs, SimCLR, and DINOv2 on downsized ISIC 2019 images.

Our work builds on these findings by systematically evaluating the interplay between resolution, computational efficiency, and diagnostic performance in the context of pretrained foundation models. By focusing on the ISIC 2019 dataset and finetuning models on downsized images, we aim to provide actionable insights for deploying AI in resource-constrained clinical environments, with potential implications for other medical imaging applications.

3. Data

We utilize the International Skin Imaging Collaboration (ISIC) 2019 dataset[17, 7, 11], comprising of 25,331 dermoscopic images labeled across eight different skin disease categories. The dataset serves as a challenging image classification dataset, as the distribution across classes is heavily skewed towards the first two classes melanoma (MEL) and melanocytic nevus (NV). The dataset is widely used for benchmarking skin lesion classification tasks due to its diversity and high-quality annotations. However, the high-resolution nature of these images (typically 1024×1024 pixels or higher) poses challenges for computational efficiency, particularly in resource-constrained settings like teledermatology platforms.

For this reason, we limit our analysis to the first two categories, MEL and NV and use a balanced dataset to create a binary classification task. To investigate the impact of image resolution on model performance, we preprocess the ISIC 2019 images by resizing them to multiple resolutions: 56×56, 112×112, and 224×224 pixels. These resolutions were selected to balance computational efficiency with the preservation of critical visual features necessary for accurate diagnosis. Additionally, we introduce simulated image degradation methods—including lossy JPEG compression, gaussian blurring, and color quantization with information loss—to mimic a wide range of transmission distortions. Each type of perturbation is randomly introduced to 20% of the dataset.

The dataset is split into training and test sets using an

80:20 ratio, ensuring that images from the same patient are not distributed across splits to prevent data leakage. This preprocessing pipeline ensures that our experiments align with the study's objectives of evaluating diagnostic performance and computational efficiency on downsized dermoscopic images, providing a robust foundation for finetuning pretrained foundation models like ViT, SimCLR, and DINOv2.

4. Methodology

4.1. Hardware

All computational tasks, including model fine-tuning and evaluation, were performed on Stanford University's Sherlock high-performance computing (HPC) cluster. Specifically, an NVIDIA H100 80GB HBM3 GPU was utilized for all processing.

4.2. Pretrained models

The core model for this study is the Vision Transformer (ViT-B/16) [10], which serves as our baseline. This specific architecture consists of a base-sized transformer with patch embeddings of size 16 × 16. The model is initialized with weights pretrained on the ImageNet dataset [8], a large-scale dataset comprising over 14 million images categorized into more than 20,000 classes. This supervised pretraining on ImageNet is a standard approach for general vision tasks, and ViT-B/16 has demonstrated significant success in a wide range of benchmarks [12]. For an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$, where H, W are height and width, and C is the number of channels, the ViT first transforms it into a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N\times(P^2\cdot C)}$, where (P, P) is the resolution of each image patch and $N = HW/P^2$ is the resulting number of patches. These patches are then linearly embedded and positional embeddings are added, forming the input sequence to the transformer encoder.

In addition to the ImageNet-pretrained ViT-B/16, we include two prominent self-supervised learning (SSL) models to evaluate whether contrastive or distillation-based pretraining enhances feature quality and robustness. These models are selected to represent distinct paradigms within SSL: one from the DINOv2 model family (a distillation-based approach) and one from SimCLR (a contrastive learning approach).

Our approach focuses on evaluating the impact of these different pretraining strategies on feature quality and robustness within the context of Vision Transformers. This is crucial for several reasons. First, by comparing supervised (ImageNet) pretraining with self-supervised methods (DINOv2, SimCLR), we aim to dissect how different learning paradigms influence the learned representations and, consequently, the downstream performance. Second, the insights gained from this study can guide future model selection and pretraining strategies for various computer vision applications, potentially leading to more efficient and robust deployments, especially in scenarios with limited labeled data. The general objective for these pretrained models is to learn a mapping $f_\theta : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{X}$ is the input image space and $\mathcal{Z}$ is the latent feature space, such that the features $\mathbf{z} = f_\theta(\mathbf{x})$ are semantically rich and discriminative.

4.2.1 DINOv2

DINOv2 [14] is a self-supervised learning framework that leverages knowledge distillation with no labels. It works by training a student Vision Transformer to match the output of a teacher Vision Transformer, where the teacher is an exponentially moving average of the student. The core idea is to learn powerful visual representations by enforcing consistency between different augmented views of the same image. For an input image $\mathbf{x}$, two augmented views $\mathbf{x}_1$ and $\mathbf{x}_2$ are generated. The student network f_s processes $\mathbf{x}_1$ and the teacher network f_t processes $\mathbf{x}_2$, producing respective representations $\mathbf{z}_s = f_s(\mathbf{x}_1)$ and $\mathbf{z}_t = f_t(\mathbf{x}_2)$. The loss function aims to minimize the discrepancy between the student's and teacher's outputs, often formulated as a cross-entropy loss or similar divergence measure:

$$\mathcal{L}_{\text{DINOv2}} = -\sum_{k=1}^{K} \mathbf{z}_{t,k} \log \mathbf{z}_{s,k}$$

where $\mathbf{z}_{t,k}$ and $\mathbf{z}_{s,k}$ are the k-th components of the teacher and student output distributions (e.g., softmax probabilities over prototypes or dimensions of the embedding). This process encourages the student to learn robust features that are invariant to various augmentations.

4.2.2 SimCLR

SimCLR (A Simple Framework for Contrastive Learning of Visual Representations) [5] is a pioneering self-supervised learning method that focuses on contrastive learning. It trains a neural network by maximizing agreement between different augmented views of the same image (positive pairs) while minimizing agreement with augmented views of other images (negative pairs). For an input image $\mathbf{x}_i$, two augmented views, $\tilde{\mathbf{x}}_{i,1}$ and $\tilde{\mathbf{x}}_{i,2}$, are generated. These views are then passed through an encoder network f to obtain representations, followed by a projection head g to project them into a latent space. The loss function, typically the NT-Xent (Normalized Temperature-scaled Cross-Entropy) loss, encourages positive pairs to be close and negative pairs to be far apart in this latent space. For a mini-batch of N images, leading to $2N$ augmented views, the loss for a positive pair $(\tilde{\mathbf{z}}_{i,1}, \tilde{\mathbf{z}}_{i,2})$ is defined as:

$$\mathcal{L}_{\text{SimCLR}} = -\log \frac{\exp(\text{sim}(\tilde{\mathbf{z}}_{i,1}, \tilde{\mathbf{z}}_{i,2})/\tau)}{\sum_{j=1}^{2N} 1_{j \neq i} \exp(\text{sim}(\tilde{\mathbf{z}}_{i,1}, \tilde{\mathbf{z}}_j)/\tau)}$$

where $\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v}/(\|\mathbf{u}\|_2 \|\mathbf{v}\|_2)$ is the cosine similarity, τ is a temperature parameter, and $1_{j \neq i}$ is an indicator function that is 1 if $j \neq i$. This framework enables the model to learn a rich representation space where semantically similar images are grouped together.

4.3. Finetuning Procedure

Model training was conducted using the Hugging Face Transformers Trainer class, providing a robust and efficient framework for finetuning Vision Transformer models. We employed the AdamW optimizer [13], a variant of Adam that incorporates decoupled weight decay, which has been shown to improve generalization performance for deep learning models. A weight decay of 0.01 was applied to regularize the model and prevent overfitting by penalizing large weights.

4.4. Initial Resolution Variance Exploration

We initiated our study by fully fine-tuning the pretrained ViT, DINOv2, and SimCLR backbones at three input resolutions: 224×224, 112×112, and 56×56. To isolate the effect of resolution, all other hyperparameters were held constant across runs: learning rate $= 1 \times 10^{-5}$, batch size $= 256$, training epochs $= 3$, and a linear learning-rate scheduler. The learning rate, epoch count, and scheduler were selected heuristically for this exploratory phase, whereas the batch size ceiling of 256 was dictated by available GPU memory.

4.5. Learning Rate Tuning

Building on the resolution sweep, we carried out a targeted learning-rate ablation. For ViT we tested three initial rates at every resolution—1×10^{-4}, 5×10^{-5}, and 1×10^{-5}. Because preliminary runs of DINOv2 at 224×224 exhibited greater loss volatility, we probed a lower range of 1×10^{-5}, 5×10^{-6}, and 1×10^{-6}.

All experiments were extended to 6 epochs and trained with a cosine-annealing scheduler that decays the learning rate smoothly from its initial value to a small floor. This schedule permits coarse exploratory updates early in training and progressively finer adjustments near convergence, a pattern that typically stabilises optimisation and improves final accuracy.

4.6. Evaluation

Model quality was assessed from three complementary perspectives:

- **Predictive accuracy.** We report top-1 accuracy for each backbone—ViT-B/16, DINOv2, and SimCLR—at all three input resolutions. This is to gauge how effectively the pretrained models can transfer their internal representations to the downstream task with varying levels of information.

- **AUC and F1-score.** To understand how well each pretraining method distinguishes between classes, especially when there's an imbalance, we used the Area Under the Receiver Operating Characteristic Curve (AUC) and the macro F1-score. While our training dataset was explicitly balanced, we were interested in exploring the accuracy per class to uncover potential disparities in performance across different categories that might not be captured by overall metrics.

- **Computational footprint.** Peak GPU memory consumption and theoretical inference FLOPs are logged for every resolution, clarifying the cost–accuracy trade-off that informs real-world deployment decisions.

5. Results

5.1. Resolution-Variant Model Performance

We first finetuned and evaluated each of the three pretrained models at multiple input resolutions. Figures 1, 2, and 3 depict the raw loss over traning steps for the ViT, DINOv2, and SimCLR finetuning experiments respectively.

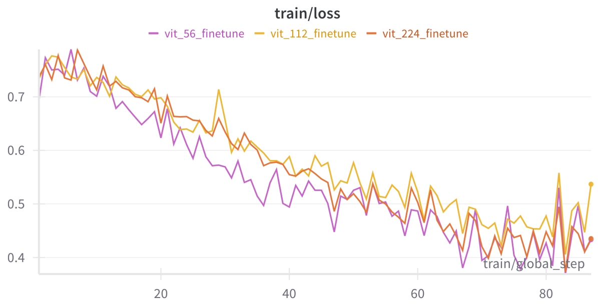

Figure 1: Training-loss trajectory for ViT at varying image resolutions

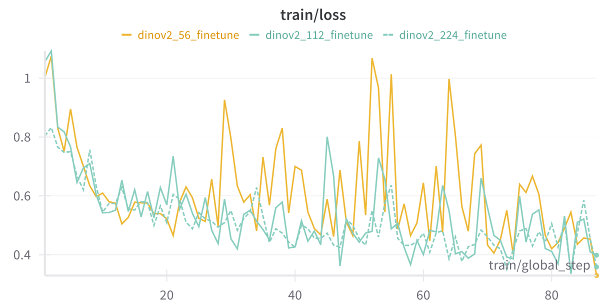

Figure 2: Training-loss trajectory for DINOv2 at varying image resolutions

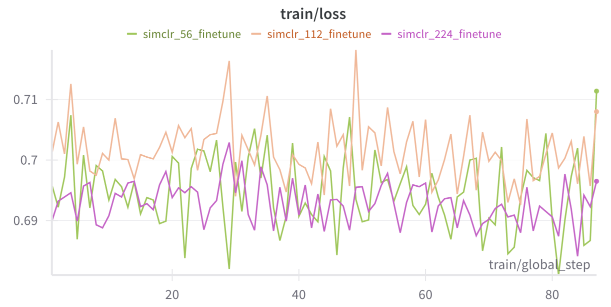

Figure 3: Training-loss trajectory for SimCLR at varying image resolutions

Higher input resolutions generally produced more favorable loss trajectories: the effect was pronounced for ViT, moderate for DINOv2, and only marginal for SimCLR.

An examination of the loss curves reveals distinct behaviors across the models. ViT's training loss showed a continuous decrease, reflecting effective learning. While DINOv2 also exhibited an overall reduction in loss, it was punctuated by significant, periodic spikes. SimCLR, however, presented a starkly different picture; its loss curves remained largely horizontal, suggesting the model struggled to learn or converge, yielding minimal gains over its training duration.

Table 1 encapsulates the core trade-offs between input resolution, accuracy, and compute:

- **Compute savings.** Shrinking images from 224 × 224 to 56 × 56 cuts FLOPs and peak GPU memory by roughly an order of magnitude for every model (e.g., ViT: 16.86 G → 0.86 G).

- **ViT resilience.** Despite the steep decrease in compute, ViT's accuracy dips marginally (0.776 → 0.752) and AUC stays ≥ 0.84.

- **DINOv2 sensitivity.** DINOv2 performs well at full resolution (0.734 accuracy, 0.879 AUC) but dramatically loses accuracy at 56 × 56, indicating stronger dependence on high-resolution detail.

- **SimCLR floor effect.** Even at 224 × 224 it lags both transformers by approximately 25 percentage points; further down-sampling yields negligible gains, confirming its limited utility under tight memory budgets.

Model	Image Resolution	Peak Memory (MB)	FLOPs (G)	Accuracy	F1	AUC
ViT	224 × 224	31 982	16.86	0.776	0.776	0.880
	112 × 112	10 002	4.28	0.756	0.755	0.847
	56 × 56	3956	0.86	0.752	0.751	0.855
DINOv2	224 × 224	46 188	21.96	0.734	0.784	0.879
	112 × 112	34 687	15.34	0.724	0.723	0.832
	56 × 56	24 567	8.97	0.688	0.630	0.819
SimCLR	224 × 224	4858	4.13	0.528	0.438	0.591
	112 × 112	2340	1.09	0.475	0.429	0.466
	56 × 56	1720	0.30	0.514	0.514	0.523

Table 1: Resource usage and validation performance across ViT, DINOv2, and SimCLR models at three input resolutions.

5.2. Learning Rate Tuning

We additionally performed learning rate tuning experiments for ViT at the three input image resolutions. Supplemental figures 4 and 5 depict the training loss trajectory and eval accuracy respectively for a 224 × 224 resolution, supplemental figures 6 and 7 depict loss and accuracy for a 112 × 112 resolution, and supplemental figures 8 and 9 depict loss and accuracy for a 56 × 56 resolution.

The learning rate experiments combined with the extended training horizon and cosine annealing produced substantial improvements in classification performance. As seen in Table 2, at a learning rate of 1×10^{-4} ViT achieved maximal gains of nearly 10 percentage points at higher resolutions (224 × 224 and 112 × 112), and over 4 percentage points at the lowest resolution (56 × 56), highlighting its sensitivity to even minimal hyperparameter tuning.

Supplementary figures 10 and 11 depict the training loss and evaluation trajectories, respectively, for DINOv2 finetuned at an input resolution of 224 × 224 across different learning rates. As shown, the revised traning setup-combined with targeted learning rate tuning yielded noticeably more stable training dynamics. Table 3 further confirms this improvement, with an over 6 percentage point increase in maximum evaluation accuracy and correspondingly improved F1 and AUC scores at a learning rate of 5×10^{-6}.

Resolution	LR	Accuracy	F1	AUC
224×224	1×10^{-4}	0.859	0.859	0.934
	5×10^{-5}	0.844	0.844	0.923
	1×10^{-5}	0.812	0.811	0.902
112×112	1×10^{-4}	0.813	0.813	0.901
	5×10^{-5}	0.802	0.805	0.896
	1×10^{-5}	0.785	0.785	0.874
56×56	1×10^{-4}	0.796	0.796	0.886
	5×10^{-5}	0.790	0.790	0.877
	1×10^{-5}	0.758	0.760	0.847

Table 2: Classification performance for ViT across input resolutions and learning rates.

LR	Accuracy	F1	AUC
1×10^{-5}	0.789	0.788	0.889
5×10^{-6}	0.812	0.797	0.902
1×10^{-6}	0.790	0.790	0.890

Table 3: Classification performance for DINOv2 at 224 × 224 resolution under three learning rates.

5.3. Results Limitations

Our experimental findings revealed a significant disparity in performance among the pretrained models when finetuned with a shared set of hyperparameters. While the chosen hyperparameters proved effective for the ImageNet-pretrained ViT-B/16, they yielded only mediocre performance for DINOv2 and notably poor results for SimCLR. We attribute these suboptimal outcomes to the inherent architectural and algorithmic differences of the self-supervised learning frameworks, which demand distinct hyperparameter configurations for optimal performance.

We hypothesize that the poor performance of SimCLR was due to the smaller than optimal batch size. SimCLR's effectiveness heavily relies on having a large number of negative samples within each batch to learn a good contrastive signal. ViTs, especially larger ones, can be memory-intensive, which can mean that it supports smaller effective batch sizes (due to memory constraints or other factors) than SimCLR would find optimal.

DINOv2's self-distillation mechanism, based on a teacher-student architecture, is inherently more stable and less prone to collapse compared to purely contrastive methods, especially when scaled. The exponential moving average (EMA) update of the teacher network provides a stable target for the student, and techniques like centering and sharpening actively prevent mode collapse without explicitly requiring a large number of negative pairs from the batch. However, despite its inherent robustness, DINOv2 also performed suboptimally with the hyperparameters tuned for ViT.

The large spikes and oscillations observed in DINOv2's training performance, despite a general decreasing trend in loss, were likely due to the fixed learning rate schedule and optimizer settings that were optimized for a different pretraining paradigm. While DINOv2 is stable, it can still benefit from a carefully tuned learning rate schedule that accounts for its distillation process, potentially requiring a slower decay or specific warm-up phase to allow the teacher to stabilize and guide the student effectively. Furthermore, the interplay between the optimizer (AdamW) and the learning rate, along with the fixed weight decay, might not have been ideally suited for DINOv2's unique loss landscape. The oscillations suggest that the model might have been frequently overshooting or oscillating around the optimal minimum, indicating a mismatch in the learning rate or momentum parameters. A dedicated hyperparameter search for DINOv2 would involve further optimizing the learning rate schedule, potentially adjusting the temperature parameters in its loss function, and exploring different momentum settings for the optimizer.

6. Discussion

This study investigated the impact of different pretraining strategies on the finetuning performance of Vision Transformers across varying image resolutions. We compared a supervised ImageNet pretraining approach (ViT-B/16) with two prominent self-supervised learning (SSL) methods, DINOv2 and SimCLR, on a downstream classification task. Our findings highlight crucial considerations regarding the transferability of learned representations and the importance of hyperparameter tuning tailored to specific pretraining paradigms.

6.1. Impact of Pretraining Strategies

The ImageNet-pretrained ViT-B/16 consistently demonstrated robust performance across all tested resolutions. This is unsurprising, given the vast scale and diversity of the ImageNet dataset, which enables the model to learn a rich set of generalizable visual features. Its strong baseline performance underscores the enduring value of large-scale supervised pretraining for many computer vision tasks. The hierarchical features learned by ViT on ImageNet appear to transfer effectively, even at lower resolutions, suggesting a degree of resolution invariance for these learned representations. Even a cursory learning-rate sweep—limited to three values—produced sizeable gains for ViT-B/16 across performance metrics. The model's sensitivity to such light-touch tuning indicates substantial head-room for further improvement through a more systematic hyperparameter search or targeted ablations. This responsiveness is particularly promising for the practical deployment of Vision

Transformers in dermatological image analysis, where task-specific data are limited and rapid iteration is essential.

DINOv2 and SimCLR exhibited more varied—and in some cases sub-optimal—performance when finetuned with the hyperparameters that worked best for ImageNet-pretrained ViT. This disparity underscores that, although self-supervised methods can learn rich representations without labels, their downstream efficacy hinges on careful task-specific tuning. Notably, even our minimal learning-rate sweep yielded non-trivial gains for DINOv2, hinting that the model may yet rival ViT once fully finetuned; a more systematic search is warranted before drawing firm conclusions. Overall, DINOv2's distillation-based training led it to outpace SimCLR, suggesting that its stability—and lack of reliance on explicit negative pairs—makes it more forgiving of sub-optimal settings. By contrast, SimCLR's performance confirmed its sensitivity to the number of negative samples, a well-documented characteristic of contrastive learning.

6.2. Resolution Variance and Efficiency

Our resolution ablation study revealed a predictable trend: higher resolutions generally correlated with improved classification accuracy across all models. This is intuitively understood as more pixel information providing richer details for discrimination. However, this gain in accuracy came at a significant computational cost. Full-resolution images, while offering the best performance, also incurred the highest inference times and memory footprints. This trade-off between performance and computational efficiency is a critical consideration for real-world applications, particularly in resource-constrained environments.

The computational profiling demonstrated that even modest reductions in resolution can yield substantial savings in inference time and memory while retaining a competitive level of accuracy. This suggests that for many practical scenarios, judicious selection of input resolution can optimize the balance between performance requirements and operational constraints. Future work could explore adaptive resolution mechanisms where the model dynamically adjusts the input resolution based on the complexity of the image or the available computational budget.

Our analysis revealed a consistent performance trend across varying model resolutions: the most effective learning rate identified for 224×224 images maintained its superiority for both 112×112 and 56×56 resolutions. We hypothesize this indicates the transferability of training strategies between different resolutions, a finding that merits further research.

6.3. Limitations and Future Directions

A primary limitation of this study was the use of a single set of hyperparameters for finetuning all pretrained models. As highlighted in the "Results Limitations" section, the sub-optimal performance of DINOv2 and especially SimCLR strongly suggests that these models require distinct hyperparameter tuning strategies for optimal transfer learning. Future work should involve comprehensive hyperparameter searches for each pretraining paradigm, including learning rate schedules, batch sizes, and potentially architecture-specific parameters (e.g., temperature in contrastive losses). This would provide a more accurate comparison of their true transfer learning capabilities.

Furthermore, while our study focused on a specific classification task, the generalizability of these findings to other downstream tasks (e.g., object detection, semantic segmentation) remains to be fully explored. Future research could investigate the transferability of these pretrained models across a wider range of computer vision benchmarks. Exploring the robustness of these models to various types of data degradation (e.g., adversarial attacks, common corruptions) would also provide valuable insights into the quality of the learned representations beyond standard accuracy metrics.

Beyond finetuning, it would be valuable to use linear probing to test the general representations learned by each model. This technique involves training a simple linear classifier on top of frozen features, providing a less biased assessment of the quality of the learned representations themselves, independent of the finetuning process.

Another promising avenue for future research is experimenting with performing further self-supervised learning for models such as DINOv2. This could involve continuing the self-supervised pretraining phase with domain-specific unlabeled data, potentially enhancing the model's ability to learn representations highly relevant to the target domain, even before finetuning with limited labeled data.

While our primary focus in this study was on general-purpose models pretrained on large natural image datasets, future extensions may compare these with domain-specific Foundation Models (FMs). This would involve evaluating models pretrained on pathology or dermatology datasets directly, assessing whether their specialized pretraining yields superior performance compared to general-purpose models adapted to the domain. Furthermore, this work can be extended to other medical domains beyond just dermatology, such as radiology, ophthalmology, or histopathology, to determine the broader applicability and transferability of these different pretraining strategies and the potential benefits of domain-specific FMs in diverse clinical settings.

7. Conclusion and Future Work

This project proposes a resource-efficient approach to dermatological diagnostics by finetuning pretrained foundation models on compressed dermoscopic images. By systematically evaluating performance across resolutions and model types, we aim to shed light on the practical trade-offs between diagnostic accuracy and computational efficiency. Our findings could catalyze more equitable access to AI-powered healthcare by enabling robust diagnostic tools in mobile and low-bandwidth environments. Moreover, the methodology and insights developed here are extensible to broader areas of medical imaging and efficient model deployment.

8. Contributions and Acknowledgments

Sonnet Xu (SX) focused on writing the experimental code, scoping the project, and writing the paper. Eric Cui (EC) focused on running the experiments, setting up compute and environments, as well as post-processing the results and writing the corresponding sections. Aaron Fanous (AF) helped to get an initial version of the code running for testing and debugging. Vicky Bikia (VB) helped with conceptualization, experimental design, and writing the paper. SX, AF, VB are all part of the Daneshjou Lab, which sponsored the compute needed for this project.

References

[1] S. Azizi, L. Culp, J. Freyberg, B. Mustafa, S. Baur, S. Kornblith, T. Chen, P. MacWilliams, S. S. Mahdavi, E. Wulczyn, B. Babenko, M. Wilson, A. Loh, P.-H. C. Chen, Y. Liu, P. Bavishi, S. M. McKinney, J. Winkens, A. G. Roy, Z. Beaver, F. Ryan, J. Krogue, M. Etemadi, U. Telang, Y. Liu, L. Peng, G. S. Corrado, D. R. Webster, D. Fleet, G. Hinton, N. Houlsby, A. Karthikesalingam, M. Norouzi, and V. Natarajan. Robust and efficient medical imaging with self-supervision, 2022.

[2] S. Azizi, B. Mustafa, F. Ryan, Z. Beaver, J. Freyberg, J. Deaton, A. Loh, A. Karthikesalingam, S. Kornblith, T. Chen, V. Natarajan, and M. Norouzi. Big self-supervised models advance medical image classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3458–3468, 2021.

[3] F. Cakir, H. Buyukdemircioglu, M. Cetin, N. Gurel, and C. Senturk. Dermatology image classification using vision transformers with limited data. *Computers in Biology and Medicine*, 148:105810, 2022.

[4] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9630–9640, 2021.

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

[6] Y.-C. Chen, X. Zhai, K. He, S. Xie, and Y. Li. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. *arXiv preprint arXiv:2206.02673*, 2022.

[7] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). *arXiv preprint arXiv:1710.05006*, 2017.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[11] C. Hernández-Pérez, M. Combalia, S. Podlipnik, N. C. F. Codella, V. Rotemberg, A. C. Halpern, O. Reiter, C. Carrera, A. Barreiro, B. Helba, S. Puig, V. Vilaplana, and J. Malvehy. BCN20000: Dermoscopic lesions in the wild. *Scientific Data*, 11(1):641, 2024.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[13] I. Loshchilov and F. Hutter. Decoupled weight decay regularization, 2019.

[14] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, V. Rivière, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[15] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in Neural Information Processing Systems*, 2019.

[16] H. Touvron, A. Vedaldi, M. Douze, and H. Jégou. Fixing the train-test resolution discrepancy. *arXiv preprint arXiv:1906.06423*, 2019.

[17] P. Tschandl, C. Rosendahl, and H. Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5:180161, 2018.

[18] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021.

9. Appendices

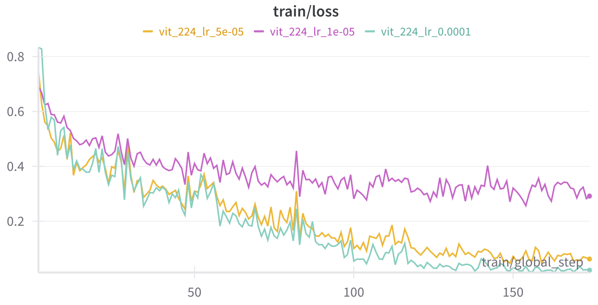

Figure 4: Training loss trajectory for ViT with 224 × 224 image resolution at varying learning rates

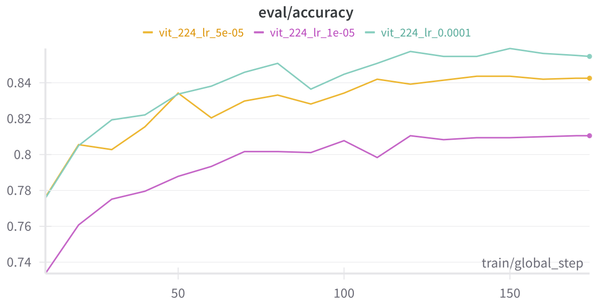

Figure 5: Eval accuracy trajectory for ViT with 224 × 224 image resolution at varying learning rates

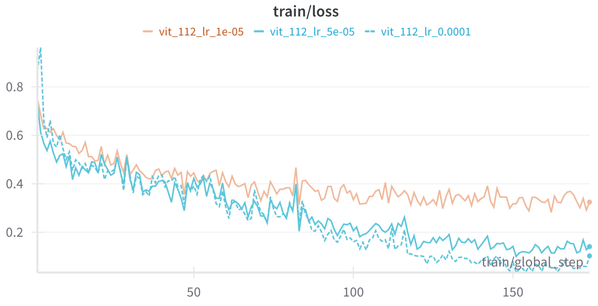

Figure 6: Training loss trajectory for ViT with 112 × 112 image resolution at varying learning rates

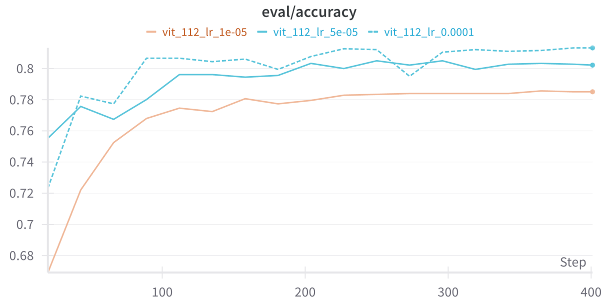

Figure 7: Eval accuracy trajectory for ViT with 112 × 112 image resolution at varying learning rates

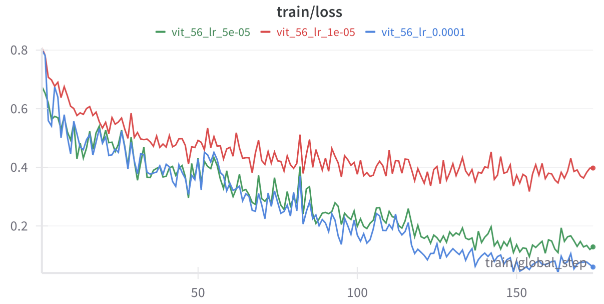

Figure 8: Training loss trajectory for ViT with 56×56 image resolution at varying learning rates

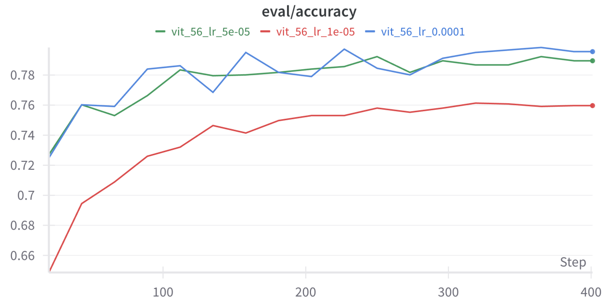

Figure 9: Eval accuracy trajectory for ViT with 56 × 56 image resolution at varying learning rates

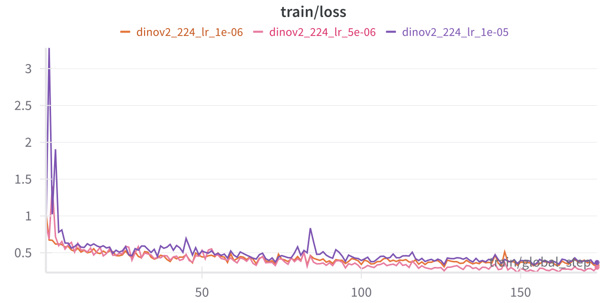

Figure 10: Training loss trajectory for DINOv2 with 224 × 224 image resolution at varying learning rates

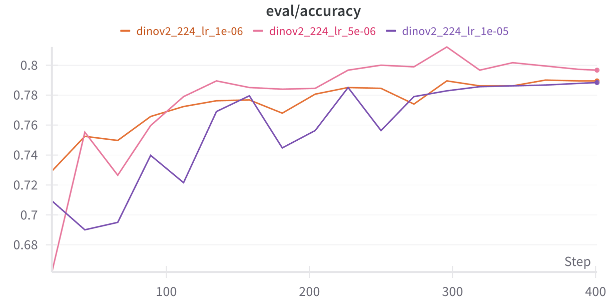

Figure 11: Eval accuracy trajectory for DINOv2 with 224 × 224 image resolution at varying learning rates