# Self-Supervised Representation Learning: A Comparative Study of MAE and SimCLR

Shivansh Pachnanda
AIMS-DTU Research Intern Selection

May 29, 2025

## Abstract

Self-supervised learning has emerged as a powerful alternative to supervised learning for visual representation learning by leveraging unlabeled data. In this report, we conduct a comparative study of two prominent self-supervised methods: Masked Autoencoders (MAE) and SimCLR.

MAE follows a masked image modeling (MIM) approach using Vision Transformers (ViT), where a high percentage of input image patches are masked, and the model learns to reconstruct the missing content. This encourages the model to understand global structure and semantics. SimCLR, on the other hand, employs contrastive learning, where the objective is to bring representations of different augmented views of the same image closer together while pushing apart representations of different images, typically using a ResNet-based encoder.

To ensure a fair and meaningful comparison, MAE ViT-Small was selected due to its comparable model size to the ResNet-50 backbone used in SimCLR. Additionally, MAE ViT-Base was included to match setups described in the original MAE paper and observe scaling effects. All models were pretrained on a subset of ImageNet-100 consisting of 25 classes and were evaluated using a linear probing setup to measure the quality of the learned representations.

The evaluation focuses on classification accuracy and provides insight into the strengths and limitations of each approach. Our results demonstrate notable differences in performance and training characteristics between MIM and contrastive learning, offering practical implications for the selection of self-supervised learning methods under resource-constrained settings. The report also discusses the challenges and trade-offs observed during training, including convergence behavior, data efficiency, and backbone architecture effects.

# Contents

# 1 Introduction

In today's data-driven era, the abundance of available data offers unprecedented opportunities to advance machine learning. Harnessing this vast information effectively is key to building robust and insightful models. However, supervised learning relies heavily on accurately labeled data, which becomes a significant bottleneck when dealing with datasets containing millions of images, a scale that is increasingly common. While labeling small amounts of data is feasible, scaling this process to massive datasets is impractical. This challenge highlights the need for methods that can learn from data in an unsupervised manner, without relying on manual annotations.

Interestingly, Wu et al. [2] observed that models trained on ImageNet using unsupervised instance discrimination were able to capture semantic similarities between classes. For example, when classifying the "panther" category, the model's second highest prediction was often "jaguar," despite not being explicitly trained with labels connecting these concepts. This demonstrates that unsupervised models can discover meaningful semantic structures in data, paving the way for powerful representation learning without labeled supervision.

Wu et al. proposed an unsupervised learning framework based on instance discrimination, where each image instance is treated as its own class. They introduced a memory bank to store feature representations of all training samples, enabling efficient comparison between the current instance and a large number of negative samples. The learning objective uses Noise Contrastive Estimation (NCE) loss, which encourages the model to distinguish between similar (positive) pairs and dissimilar (negative) pairs. Based on this approach, the learned embeddings stored in the memory bank can be used with weighted k-Nearest Neighbors (k-NN) to identify closest pairs, facilitating pseudo-labeling.

Self-Supervised Learning (SSL) represents a broader category of approaches that aim to learn feature representations without manual labels. Instead of relying on annotations, SSL methods create proxy or pretext tasks that enable the model to learn useful patterns from the inherent structure of the data itself, for example, predicting missing parts of an image or distinguishing between different augmented views of the same image.

The key idea behind SSL is to pretrain large models on massive unlabeled datasets, enabling them to learn generalized and transferable features. These pretrained models can then be fine-tuned or evaluated on smaller labeled datasets for specific downstream tasks. This pretraining-finetuning paradigm allows for improved performance and efficiency, especially in scenarios where labeled data is scarce or expensive to obtain.

# 2 Background

## 2.1 Masked Autoencoders (MAE)

Masked Autoencoders (MAE), introduced by He et al. [?], are a self-supervised learning approach designed to learn powerful visual representations through masked image modeling (MIM). Inspired by the success of masked language modeling in natural language processing (e.g., BERT), MAE applies a similar concept to images: it randomly masks a large portion (typically 75

At its core, MAE employs an asymmetric encoder-decoder architecture composed of two Vision Transformers. The encoder receives only the visible patches, about 25

ViTs are known to be highly data-hungry and require large-scale pretraining to perform well. MAE's masked reconstruction task acts as an effective pretraining objective, allowing the ViT encoder to learn meaningful image features even when trained on unlabeled data. These pretrained encoders can then be fine-tuned on downstream tasks such as image classification, segmentation, and object detection with relatively smaller labeled datasets. This pretrain-finetune paradigm addresses the challenge of training data scarcity while leveraging the model's high capacity.

The input image is first divided into fixed-size patches which are flattened and linearly projected into patch embeddings. These embeddings are randomly shuffled, and only a subset corresponding to visible patches is passed to the encoder. After encoding, the embeddings are restored to their original order, mask tokens are added for the missing patches, and the decoder reconstructs the full image. By forcing the model to predict masked pixels based on visible context, MAE encourages learning of deep, high-level representations that generalize well across tasks.

MAE's simplicity, scalability, and efficiency make it a promising approach to self-supervised learning in computer vision, especially for leveraging the power of Vision Transformers without requiring massive labeled datasets.

### 2.2 SimCLR

SimCLR, introduced by Chen et al. [4], is a popular self-supervised learning framework based on contrastive learning, which learns visual representations by maximizing agreement between differently augmented views of the same image. Unlike methods relying on memory banks or complex negative sampling strategies, SimCLR simplifies the pipeline by using large batch sizes to provide a rich set of negative samples within each batch.

The key idea in SimCLR is to generate two distinct augmented versions (or "views") of each input image through random transformations such as cropping, color distortion, and Gaussian blur. These augmented pairs are then passed through a shared encoder network, typically a ResNet, which outputs feature embeddings. A small neural network called the projection head maps these embeddings into a space where contrastive loss is applied.

The loss used, known as the normalized temperature-scaled cross entropy loss (NT-Xent), encourages the model to bring embeddings of positive pairs (augmentations of the same image) closer while pushing apart embeddings of negative pairs (different images in the batch). This is achieved without requiring explicit labels, enabling the model to learn meaningful representations from unlabeled data.

One of SimCLR's important design choices is the use of very large batch sizes during pretraining. Large batches provide many negative samples, making the contrastive task more effective without needing a separate memory bank to store features, unlike earlier approaches such as *Wu et al.*. This simplification reduces implementation complexity and leverages modern hardware capabilities.

After pretraining, the learned encoder can be evaluated by freezing its weights and training a simple linear classifier on top of the embeddings (linear probing) to assess representation quality. This demonstrates that the learned features are general and useful for downstream tasks such as image classification, even without task-specific finetuning.

SimCLR's approach highlights the effectiveness of contrastive learning with careful data augmentation and large batch training, making it a strong baseline for self-supervised visual representation learning.

## 3 Methodology

### 3.1 Dataset

We utilized the ImageNet-100 dataset for this study, which consists of 100 classes with colored images of varying sizes. The training set was divided into four groups, each containing 25 classes with approximately 1300 images per class, while the validation set included all 100 classes with 50 images per class. To improve model generalization and simulate different views of the same image, data augmentations were applied during training. The augmentations differed between SimCLR and MAE training pipelines, reflecting their respective methodologies.

These augmentations were based on the recommendations from the original papers and differed slightly between SimCLR and MAE training pipelines.

**SimCLR Augmentations:**

- Random resized crop
- Random horizontal flip
- Color jittering (brightness, contrast, saturation, hue)
- Random grayscale conversion
- Gaussian blur

**MAE Augmentations:**

- Random resized crop with scale range (0.2, 1.0) using bicubic interpolation
- Random horizontal flip
- Normalization using ImageNet mean and standard deviation

### 3.2 Pretraining Setup

All models were pretrained on Kaggle using two NVIDIA T4 GPUs with a data-parallel pipeline to leverage multi-GPU acceleration. In SimCLR, batch normalization (BN) can inadvertently leak information across samples within a batch, causing the model to converge faster but potentially learn poorer quality representations. This phenomenon necessitates large batch sizes and careful handling of BN statistics. To mitigate this, we used a multi-GPU setup to maintain synchronized BN across devices. While alternatives like replacing BN with Layer Normalization have been proposed, our experiments relied on multiple GPUs to address this challenge.

For SimCLR, a batch size of 256 was used, while for both MAE-Base and MAE-Small models, a batch size of 512 was maintained. Each model was trained for 100 epochs. The data loaders utilized 4 CPU workers for efficient data preprocessing, consistent with the 4-core CPU environment available on Kaggle. Gradient accumulation was set to 1, indicating no accumulation was performed.

### 3.3 Linear Probing Setup

To evaluate the quality of the representations learned by the self-supervised models, we performed linear probing. Instead of fine-tuning the entire model, which is computationally expensive, we treated the pretrained self-supervised encoders as frozen feature extractors. A lightweight classifier was then trained on top using cross-entropy loss, making the evaluation process efficient while still indicative of representation quality.

All 100 classes from the ImageNet-100 validation set were included for this downstream task. The probing classifier consisted of a short `nn.Sequential` block, beginning with a `BatchNorm1d` layer for normalization, followed by a single linear layer outputting logits for 100 classes. This normalization trick, although not discussed in the original SimCLR or MAE papers, is commonly used in practice and was mentioned in the official reference implementations of MAE. It has been found to improve classification accuracy by 1–2%.

We used a batch size of 512 during linear probing, and trained the classifier for 40 epochs per checkpoint. Checkpoints from the SSL pretraining phase were sampled every 5 epochs (i.e., at epochs 5, 10, 15, ..., 100), and the probing classifier was trained on each to track the evolution of feature quality throughout training.

Evaluation results from this probing phase are presented in the form of graphs later in the report.

## 4 Experimental Results

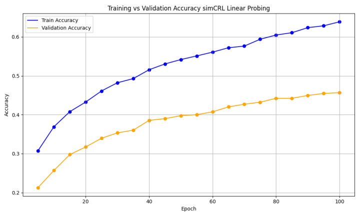

Figure 1: SimCLR

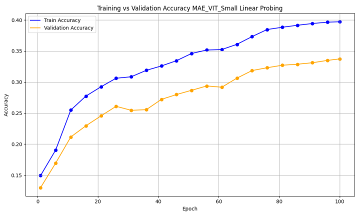

Figure 2: MAE Small

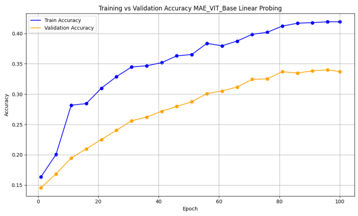

Figure 3: MAE Base

| Epoch | Loss | Accuracy | Val Accuracy |
|---|---|---|---|
| 1 | 3.7521 | 0.1493 | 0.1298 |
| 6 | 3.5340 | 0.1904 | 0.1694 |
| 11 | 3.2107 | 0.2548 | 0.2114 |
| 16 | 3.0970 | 0.2773 | 0.2294 |
| 21 | 3.0139 | 0.2923 | 0.2454 |
| 26 | 2.9376 | 0.3060 | 0.2608 |
| 31 | 2.9185 | 0.3084 | 0.2544 |
| 36 | 2.8712 | 0.3188 | 0.2554 |
| 41 | 2.8243 | 0.3258 | 0.2720 |
| 46 | 2.7822 | 0.3343 | 0.2798 |
| 51 | 2.7257 | 0.3460 | 0.2866 |
| 56 | 2.6952 | 0.3514 | 0.2936 |
| 61 | 2.6879 | 0.3523 | 0.2916 |
| 66 | 2.6448 | 0.3607 | 0.3062 |
| 71 | 2.5779 | 0.3731 | 0.3184 |
| 76 | 2.5241 | 0.3845 | 0.3230 |
| 81 | 2.5127 | 0.3882 | 0.3270 |
| 86 | 2.4881 | 0.3914 | 0.3286 |
| 91 | 2.4793 | 0.3942 | 0.3310 |
| 96 | 2.4683 | 0.3964 | 0.3348 |
| 100 | 2.4674 | 0.3970 | 0.3372 |

Table 1: Linear Probing Results for MAE VIT Small

| Epoch | Loss | Accuracy | Val Accuracy |
|---|---|---|---|
| 1 | 3.6800 | 0.1637 | 0.1458 |
| 6 | 3.4884 | 0.2008 | 0.1686 |
| 11 | 3.0780 | 0.2818 | 0.1948 |
| 16 | 3.0471 | 0.2844 | 0.2098 |
| 21 | 2.9237 | 0.3099 | 0.2252 |
| 26 | 2.8249 | 0.3288 | 0.2404 |
| 31 | 2.7420 | 0.3448 | 0.2560 |
| 36 | 2.7375 | 0.3468 | 0.2622 |
| 41 | 2.6974 | 0.3519 | 0.2718 |
| 46 | 2.6487 | 0.3632 | 0.2798 |
| 51 | 2.6399 | 0.3654 | 0.2874 |
| 56 | 2.5480 | 0.3838 | 0.3010 |
| 61 | 2.5672 | 0.3797 | 0.3052 |
| 66 | 2.5307 | 0.3875 | 0.3118 |
| 71 | 2.4815 | 0.3987 | 0.3244 |
| 76 | 2.4673 | 0.4020 | 0.3252 |
| 81 | 2.4162 | 0.4123 | 0.3370 |
| 86 | 2.3942 | 0.4170 | 0.3348 |
| 91 | 2.3909 | 0.4180 | 0.3384 |
| 96 | 2.3846 | 0.4193 | 0.3400 |
| 100 | 2.3834 | 0.4193 | 0.3374 |

Table 2: Linear Probing Results for MAE VIT Base

# 5 Discussion

## 5.1 Limitations

- The entire pretraining and evaluation process was constrained to a total of approximately 60 GPU hours. Notably, SimCLR pretraining alone consumed around 20 hours, which is

| Epoch | Loss | Accuracy | Val Accuracy |
|---|---|---|---|
| 5 | 2.8985 | 0.3077 | 0.2128 |
| 10 | 2.6041 | 0.3690 | 0.2568 |
| 15 | 2.4149 | 0.4078 | 0.2976 |
| 20 | 2.2978 | 0.4325 | 0.3172 |
| 25 | 2.1750 | 0.4610 | 0.3396 |
| 30 | 2.0802 | 0.4821 | 0.3534 |
| 35 | 2.0328 | 0.4927 | 0.3604 |
| 40 | 1.9308 | 0.5155 | 0.3856 |
| 45 | 1.8654 | 0.5305 | 0.3900 |
| 50 | 1.8211 | 0.5416 | 0.3976 |
| 55 | 1.7811 | 0.5510 | 0.4002 |
| 60 | 1.7346 | 0.5607 | 0.4074 |
| 65 | 1.6842 | 0.5720 | 0.4202 |
| 70 | 1.6710 | 0.5763 | 0.4270 |
| 75 | 1.5938 | 0.5938 | 0.4322 |
| 80 | 1.5488 | 0.6046 | 0.4422 |
| 85 | 1.5255 | 0.6106 | 0.4422 |
| 90 | 1.4765 | 0.6236 | 0.4496 |
| 95 | 1.4454 | 0.6284 | 0.4546 |
| 100 | 1.4108 | 0.6386 | 0.4568 |

Table 3: Linear Probing Results for SimCRL

| **Model** | **Architecture** | **Params (M)** | **Time/Epoch** | **Train Acc.** | **Val Acc.** |
|---|---|---|---|---|---|
| MAE VIT Small | Vision Transformer (ViT Small) | 21.7M | 2 min 57 sec | 39.7% | 33.7% |
| MAE VIT Base | Vision Transformer (ViT Base) | 86.6M | 4 min 34 sec | 41.9% | 34.0% |
| SimCLR | ResNet-50 | 23.5M | 12 min 02 sec | 63.9% | 45.7% |

Table 4: Comparison of MAE and SimCLR models in terms of architecture, parameter count, training time, and linear probing accuracy

significant given the limited compute budget.

- Due to restricted compute resources and time, hyperparameter tuning and extensive experimentation, especially for MAE variants, were limited, potentially affecting the optimal performance achievable by these models.
- This study represents my first experience working on self-supervised learning tasks and large-scale visual representation learning, which impacted the depth of experimentation and optimization.

### 5.2 Observations

- **Overfitting in SimCLR:** A substantial gap is observed between the training and validation accuracy for SimCLR, suggesting a potential overfitting issue. This is likely due to the use of Batch Normalization (BN) in settings with limited batch diversity, especially when training on a single GPU. In future iterations, shuffling batches across devices (as done in MoCo [3]) or increasing the number of GPUs to improve BN statistics could help mitigate this problem.
- **Contrastive Learning is Better Suited for Linear Probing:** SimCLR demonstrates superior performance in linear probing compared to MAE-based methods. This aligns with findings in the literature, where contrastive methods tend to learn representations that are more linearly separable. The global objective of maximizing agreement between positive pairs lends itself well to downstream evaluation using linear classifiers.

- **MAE Performance is Typically Realized Through Fine-tuning:** Masked Image Modeling (MIM) techniques like MAE are not generally optimized for evaluation via linear probing. Instead, they excel when followed by full fine-tuning on the downstream task. The relatively poor linear probing results for MAE models, despite their strong architecture and training efficiency, reflect this characteristic.
- **Impact of Dataset Scale:** The overall accuracy values for all models are quite low. This is likely due to the relatively small scale of the dataset used for self-supervised pretraining. Both MAE and SimCLR require substantial data to learn generalizable visual representations. In practical scenarios, such models are often pretrained on millions of images (e.g., ImageNet or larger datasets).
- **Efficiency vs. Performance Trade-off:** MAE ViT Small was the most efficient model in terms of training time per epoch (2 min 57 sec), but SimCLR achieved significantly higher linear probing accuracy at the cost of much higher training time (12 min 02 sec). This highlights a trade-off between computational efficiency and representational quality in self-supervised learning methods.
- **Architecture Biases:** ViT architectures used in MAE models may be less suited to small datasets and simple classification heads without extensive fine-tuning or strong inductive biases (e.g., convolutional priors). This might explain the relatively lower performance of MAE compared to SimCLR, which uses a ResNet backbone with strong spatial biases and regularization properties.

## 6 Conclusion

In this study, we conducted a comparative evaluation of self-supervised learning models, MAE (ViT Small and Base) and SimCLR (ResNet-50), using linear probing as the downstream task. The results highlight several key findings:

- SimCLR significantly outperforms MAE models in terms of linear probing accuracy, suggesting that contrastive learning methods are more aligned with this evaluation protocol.
- MAE models, although efficient in terms of training time and parameter count (especially ViT Small), showed limited performance when not fine-tuned, which aligns with the broader consensus that Masked Image Modeling (MIM) methods are better suited for full fine-tuning.
- The observed overfitting in SimCLR's results indicates the importance of batch normalization handling and the need for multi-GPU strategies, such as those used in MoCo, to enhance generalization.
- Across all models, the limited dataset size was a major bottleneck. Larger datasets are essential to fully leverage the potential of self-supervised learning methods.

Future work will focus on fine-tuning MAE-based models, exploring stronger augmentations and longer training schedules for SimCLR, and scaling to more realistic data sizes to better assess generalizability and robustness.

## References

[1] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, Ross Girshick. *Masked Autoencoders Are Scalable Vision Learners.* CVPR 2022.
`https://github.com/facebookresearch/mae`

[2] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, Dahua Lin. *Unsupervised Feature Learning via Non-Parametric Instance Discrimination.* CVPR 2018.

[3] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross Girshick. *Momentum Contrast for Unsupervised Visual Representation Learning.* CVPR 2020.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton. *A Simple Framework for Contrastive Learning of Visual Representations.* ICML 2020.
`https://github.com/sthalles/SimCLR`

# 7 Appendix

All code available on Github `https://github.com/KahnSvaer/Comparitive_Analysis_SSL_tasks_MIM_ID`