

# Benchmarking Self-Supervised Learning on STL-10: SimCLR vs BYOL

## Abstract

Self-supervised learning (SSL) has emerged as a powerful paradigm for learning visual representations without relying on large amounts of labeled data. In this work, we conduct a systematic benchmark of two prominent SSL methods—SimCLR and BYOL—on the STL-10 dataset. We evaluate the quality of learned representations using standard downstream protocols: linear probing and k-nearest neighbor (k-NN) classification. Our results show that BYOL consistently outperforms SimCLR under identical experimental settings, achieving a linear probe accuracy of 70.14% compared to 63.06% for SimCLR, and a k-NN accuracy of 68.66% compared to 58.66%. We additionally discuss practical constraints encountered when attempting to scale the benchmark to transformer-based SSL methods such as DINO, and justify their exclusion due to hardware limitations. This study provides a clean, reproducible comparison suitable for understanding SSL behavior under limited computational budgets.

---

## 1. Introduction

Deep learning models for computer vision traditionally rely on large, labeled datasets such as ImageNet. However, collecting labeled data is expensive, time-consuming, and often infeasible for specialized domains. Self-supervised learning addresses this issue by learning useful representations from unlabeled data through carefully designed pretext tasks or objectives.

Contrastive and non-contrastive SSL methods have recently demonstrated that models trained without labels can rival or even surpass supervised pretraining when evaluated on downstream tasks. Among these, SimCLR introduced a simple yet effective contrastive framework, while BYOL proposed a non-contrastive alternative that avoids the need for explicit negative samples.

The goal of this project is not to achieve state-of-the-art performance, but to perform a controlled, research-grade comparison between SimCLR and BYOL under identical conditions. We focus on STL-10, a medium-scale dataset commonly used in SSL research, and evaluate representation quality using well-established protocols.

---

## 2. Background and Related Work

### 2.1 Self-Supervised Learning

Self-supervised learning methods learn representations by creating surrogate supervision signals from the data itself. In vision, this typically involves generating multiple views of the same image through data augmentation and enforcing consistency between their representations.

### 2.2 SimCLR

SimCLR is a contrastive learning framework that relies on instance discrimination. Given two augmented views of the same image (a positive pair), the model is trained to maximize their similarity while minimizing similarity with other images in the batch (negative samples). Key components of SimCLR include:

- Strong data augmentations
- A projection head (MLP) applied on top of the encoder
- A contrastive loss (NT-Xent) with temperature scaling

SimCLR's performance is known to be sensitive to batch size, as larger batches provide more negative samples.

### 2.3 BYOL

Bootstrap Your Own Latent (BYOL) removes the need for negative samples entirely. It uses two networks:

- A student network, trained via gradient descent
- A teacher network, updated as an exponential moving average (EMA) of the student

The student is trained to predict the teacher's representation of another augmented view of the same image. Despite lacking explicit collapse-prevention mechanisms such as negative samples, BYOL empirically avoids trivial solutions when properly configured.

---

## 3. Dataset

### 3.1 STL-10

STL-10 consists of:

- 100,000 unlabeled images
- 5,000 labeled training images
- 8,000 labeled test images
- 10 object categories

Images are of size  $96 \times 96$ , making STL-10 suitable for experimentation under limited computational resources while still being large enough to demonstrate meaningful SSL behavior.

In this benchmark:

- The unlabeled split is used for self-supervised pretraining
  - The labeled train and test splits are used exclusively for evaluation
- 

## 4. Methodology

### 4.1 Encoder Architecture

Both SimCLR and BYOL use ResNet-18 as the backbone encoder. The final classification layer is removed, and the output feature dimension is 512. This choice ensures a fair comparison and keeps computational requirements manageable.

### 4.2 Data Augmentations

Consistent with prior work, strong stochastic augmentations are applied to generate multiple views of each image:

- Random resized cropping
- Horizontal flipping
- Color jitter
- Random grayscale
- Gaussian blur

For SimCLR, two augmented views are generated per image. For BYOL, two views are also used, but they are processed by different network branches (student and teacher).

## 4.3 Training Details

Both models are trained using:

- Adam optimizer
- Cosine learning rate scheduling
- Mixed precision training (AMP)

Care was taken to ensure that both methods were trained for sufficient epochs to reach convergence. BYOL was observed to converge faster and more stably than SimCLR under identical batch sizes.

---

## 5. Evaluation Protocols

### 5.1 Linear Probing

Linear probing is the primary evaluation protocol used in SSL research. The procedure is as follows:

1. Freeze the pretrained encoder
2. Train a single linear classifier on top of the frozen features using labeled training data
3. Evaluate classification accuracy on the test set

High linear probe accuracy indicates that the learned representations are linearly separable and semantically meaningful.

### 5.2 k-Nearest Neighbor (k-NN) Evaluation

k-NN evaluation measures the intrinsic quality of the representation space without training any classifier:

1. Extract normalized features for all training and test images
2. For each test image, find its k nearest neighbors in feature space
3. Predict the label via majority voting

We use  $k = 20$ , a commonly adopted value in SSL benchmarks.

---

## 6. Results

### 6.1 Quantitative Results

Model	Linear Probe Accuracy	k-NN Accuracy (k=20)
SimCLR	63.06%	58.66%
BYOL	<b>70.14%</b>	<b>68.66%</b>

### 6.2 Analysis

BYOL outperforms SimCLR by a significant margin across both evaluation metrics. The performance gap is particularly pronounced in k-NN evaluation, suggesting that BYOL learns a more structured and semantically coherent feature space.

SimCLR's reliance on negative samples and large batch sizes likely limits its effectiveness under constrained compute settings. BYOL, by contrast, demonstrates strong robustness even with moderate batch sizes, aligning with findings reported in the original literature.

---

## 7. Attempted Extension: DINO

### 7.1 Motivation

DINO is a self-distillation method that applies SSL to Vision Transformers (ViTs) and has shown exceptional results on large-scale datasets. Including DINO would have extended this benchmark to transformer-based SSL.

### 7.2 Practical Constraints

Despite multiple attempts, DINO could not be reliably trained within the available hardware constraints:

- Limited GPU memory on the local machine
- Restricted VRAM and runtime quotas on the Kaggle free tier
- High computational cost due to multi-crop augmentations and ViT backbones

Training repeatedly resulted in out-of-memory errors, unstable convergence, or prohibitively slow runtimes. While various optimizations were explored (reduced crops, smaller heads, mixed

precision), these compromises significantly deviated from the standard DINO setup and risked invalidating comparisons.

### 7.3 Decision Rationale

Rather than including an underpowered or unstable DINO experiment, we chose to focus on producing a clean, rigorous comparison between SimCLR and BYOL. This decision prioritizes experimental validity and reproducibility over breadth.

---

## 8. Discussion

This benchmark highlights several important insights:

- Non-contrastive methods such as BYOL can outperform contrastive approaches under limited compute
- Representation quality should be evaluated using multiple protocols
- Practical constraints play a crucial role in method selection, especially outside large research labs

The results reinforce the idea that careful experimental design and evaluation are more valuable than simply adopting the most complex or computationally demanding methods.

---

## 9. Conclusion

In this work, we presented a controlled benchmark of SimCLR and BYOL on the STL-10 dataset. BYOL consistently demonstrated superior representation learning, achieving higher linear probe and k-NN accuracies. While transformer-based SSL methods such as DINO are promising, their computational demands make them impractical under constrained environments.

Overall, this project demonstrates that meaningful self-supervised learning research can be conducted with modest resources, provided that experiments are carefully designed and evaluated.

---

## 10. Future Work

Future extensions of this work could include:

- Augmentation ablation studies
  - Evaluation on additional datasets
  - Semi-supervised fine-tuning
  - Scaling experiments under access to higher-end hardware
- 

## Acknowledgements

This work builds upon foundational ideas introduced by Chen et al. (SimCLR) and Grill et al. (BYOL), and aims to serve as a practical, reproducible study for understanding self-supervised representation learning.