CSE475 - Self-Supervised Learning
Rice Variety Classification

Department of Computer Science and Engineering
East West University, Dhaka, Bangladesh

October 29, 2025

Abstract

This document presents the selection and detailed explanation of three state-of-the-art self-supervised learning (SSL) methods for fine-grained rice variety classification: **SimCLR**, **BYOL**, and **DINO**. These methods are strategically chosen based on their proven effectiveness in fine-grained visual classification tasks and their suitability for our dataset characteristics (38 rice varieties, 19,000 images with subtle inter-class differences). Each method represents a different learning paradigm: contrastive learning (SimCLR), predictive learning without negatives (BYOL), and knowledge distillation with multi-crop strategy (DINO). The comprehensive analysis includes mathematical formulations, architectural details, implementation considerations, and expected performance improvements over supervised baselines.

1 Introduction

1.1 Project Context

Fine-grained visual classification remains one of the most challenging problems in computer vision, particularly when dealing with agricultural products where subtle morphological differences distinguish between categories. Our rice variety classification task exemplifies this challenge with the following characteristics:

- **Task:** Fine-grained classification of 38 rice varieties from Bangladesh
- **Dataset Size:** 19,000 images (500 per class)
- **Image Resolution:** 640×480 pixels (resized to 224×224 for training)
- **Key Challenges:**
  - Subtle morphological differences between varieties
  - Variations in imaging conditions (lighting, angle, background)
  - High intra-class variation and inter-class similarity
  - Need for robust feature extraction mechanisms

From Task 2, our supervised baseline models achieved approximately 71-75% accuracy across various train-test splits, indicating substantial room for improvement through self-supervised pre-training approaches.

1.2 Motivation for Self-Supervised Learning

Self-supervised learning offers compelling advantages for our rice classification task:

1. **Robust Representation Learning:** SSL methods learn meaningful visual representations from unlabeled data, potentially capturing features that supervised learning might miss
2. **Fine-grained Feature Discovery:** The pretext tasks in SSL can help models focus on discriminative features critical for distinguishing similar rice varieties
3. **Improved Generalization:** By learning invariant representations, SSL models often generalize better to unseen data
4. **Data Efficiency:** SSL can reduce dependence on large labeled datasets, particularly valuable in agricultural domains where expert labeling is expensive

2 Selected SSL Methods

2.1 Selection Criteria

Our three methods were selected based on the following criteria:

1. **Proven Effectiveness:** State-of-the-art performance on fine-grained classification benchmarks

2. **Computational Feasibility:** Trainable within Kaggle's GPU constraints (P100/T4)
3. **Complementary Approaches:** Different learning paradigms for comprehensive comparison
4. **Dataset Suitability:** Appropriate for texture-based and morphological classification tasks
5. **Implementation Maturity:** Well-documented methods with available implementations

3 Method 1: SimCLR (Simple Framework for Contrastive Learning)

3.1 Intuition and Core Principle

SimCLR [1] operates on the principle that different augmented views of the same image should produce similar representations, while views from different images should be distinct. For rice grains, this translates to: *"Different views of the same rice grain should have similar representations, regardless of transformations applied."*

The method learns by maximizing agreement between positive pairs (different augmentations of the same image) while minimizing similarity to negative pairs (augmentations of different images).

3.2 Architecture and Components

The SimCLR framework consists of four main components:

1. **Data Augmentation Module:** Applies stochastic transformations to create multiple views
2. **Base Encoder:** CNN backbone (ResNet50/EfficientNet) for feature extraction
3. **Projection Head:** Multi-layer perceptron mapping features to contrastive space
4. **Contrastive Loss Function:** NT-Xent loss maximizing positive pair agreement

3.2.1 Mathematical Formulation

The NT-Xent (Normalized Temperature-scaled Cross Entropy) loss is defined as:

$$\mathcal{L}_{i,j} = -\log\left[\frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbf{1}_{[k\neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}\right] \tag{1}$$

where:

- $\mathbf{z}_i, \mathbf{z}_j$ are projections of positive pair (same image, different augmentations)
- $\mathbf{z}_k$ are projections of all other images in the batch (negatives)
- $\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T\mathbf{v}/(||\mathbf{u}||||\mathbf{v}||)$ is cosine similarity
- τ is the temperature parameter (typically 0.5)
- N is the batch size, $2N$ is total augmented samples

3.2.2 Data Augmentation Strategy

For rice grain images, we employ the following augmentation pipeline:

- Random resized crop (scale: 0.08 to 1.0)
- Random horizontal flip (probability: 0.5)
- Color jitter (brightness: ±0.4, contrast: ±0.4, saturation: ±0.4, hue: ±0.1)
- Gaussian blur ($\sigma \in [0.1, 2.0]$, probability: 0.5)
- Random grayscale conversion (probability: 0.2)

3.3 Training Procedure

Hyperparameters:

- Batch Size: 256 (critical for diverse negative samples)
- Temperature: $\tau = 0.5$
- Optimizer: LARS with cosine learning rate decay
- Base Learning Rate: 0.3
- Training Epochs: 200-300
- Weight Decay: 1×10^{-6}

3.4 Suitability for Rice Classification

SimCLR offers several advantages for our task:

1. **Robustness to Imaging Variations:** The strong augmentation pipeline teaches the model to focus on intrinsic grain characteristics rather than imaging artifacts
2. **Texture and Color Learning:** Contrastive learning excels at capturing subtle texture patterns and color variations critical for rice variety identification
3. **Large Negative Set:** With 38 similar classes, large batch sizes provide diverse negative samples, preventing spurious correlations
4. **Proven Fine-grained Performance:** SimCLR has demonstrated strong results on fine-grained datasets like CUB-200 and Stanford Cars

Expected Performance: +2-4% improvement over supervised baseline, requiring 200-250 epochs for convergence.

4 Method 2: BYOL (Bootstrap Your Own Latent)

4.1 Intuition and Core Principle

BYOL [2] eliminates the need for negative pairs by using a momentum-based target network. The online network learns to predict the representation that a slow-moving target network would produce for a different augmentation of the same image. The key insight: *"Predict how a stable teacher network would represent a different augmentation of the same rice grain."*

4.2 Architecture and Components

BYOL employs a dual-network architecture:

1. **Online Network (θ):** Actively learning encoder with predictor
2. **Target Network (ξ):** Exponential moving average of online network

4.2.1 Mathematical Formulation

The prediction loss is defined as:

$$\mathcal{L}_{\theta,\xi} = \frac{||\mathbf{q}_\theta(\mathbf{z}_\theta) - \text{sg}(\mathbf{z}'_\xi)||_2^2}{||\mathbf{q}_\theta(\mathbf{z}_\theta)||_2 \cdot ||\mathbf{z}'_\xi||_2} \tag{2}$$

where:

- $\mathbf{q}_\theta(\mathbf{z}_\theta)$ is the online network prediction (normalized)
- $\mathbf{z}'_\xi$ is the target network projection (normalized)
- $\text{sg}(\cdot)$ is the stop-gradient operator

The target network is updated via exponential moving average:

$$\xi \leftarrow \tau\xi + (1-\tau)\theta \tag{3}$$

where $\tau = 0.996$ is the momentum coefficient.

4.3 Training Procedure

Hyperparameters:

- Batch Size: 64-128 (much smaller than SimCLR)
- Momentum: $\tau = 0.996$
- Optimizer: AdamW
- Learning Rate: 3×10^{-4} with cosine decay
- Training Epochs: 300-400
- Weight Decay: 1×10^{-5}

4.4 Advantages for Rice Classification

BYOL provides several benefits for our fine-grained task:

1. **No Negative Pairs:** Eliminates potential confusion between similar rice varieties
2. **GPU Efficiency:** Smaller batch sizes reduce memory requirements
3. **Training Stability:** Momentum update mechanism provides stable learning targets
4. **Fine-grained Effectiveness:** Predictor network learns nuanced transformations between views

Expected Performance: +3-5% improvement over supervised baseline, with stable convergence in 300-350 epochs.

5 Method 3: DINO (Self-Distillation with No Labels)

5.1 Intuition and Core Principle

DINO [3] combines self-supervised learning with knowledge distillation using a multi-crop strategy and Vision Transformers. A student network learns from a momentum-updated teacher network by matching output distributions across multiple crops. The key innovation: *"Student learns global grain characteristics while discovering fine-grained texture patterns through local crops."*

5.2 Architecture and Multi-Crop Strategy

DINO's multi-crop strategy enables multi-scale learning:

- **Global Crops (2 views):** 224×224 resolution, capture overall grain shape and structure
- **Local Crops (6-8 views):** 96×96 resolution, capture fine-grained texture details

5.2.1 Mathematical Formulation

The cross-entropy loss with centering is:

$$\mathcal{L} = -\sum_{x \in \{x_g^1, x_g^2\}} \sum_{x' \in \mathcal{X}} P_t(x) \log P_s(x') \tag{4}$$

where $\{x_g^1, x_g^2\}$ are global crops and $\mathcal{X}$ is the set of all crops.

The teacher output with centering and sharpening:

$$P_t(x)^{(i)} = \frac{\exp((g_t(x)^{(i)} - c^{(i)})/\tau_t)}{\sum_j \exp((g_t(x)^{(j)} - c^{(j)})/\tau_t)} \tag{5}$$

where c is the center vector (running average) and τ_t is the teacher temperature.

5.3 Vision Transformer Architecture

DINO leverages Vision Transformers (ViT) for superior fine-grained understanding:

- **Self-Attention:** Captures long-range dependencies within grain regions
- **Attention Maps:** Provide interpretability for discriminative features
- **Patch-Based Processing:** Natural fit for grain images with distinct regions

5.4 Training Procedure

Hyperparameters:

- Batch Size: 64-128
- Global Crops: 2 views at 224×224
- Local Crops: 8 views at 96×96
- Teacher Temperature: $\tau_t = 0.04$
- Student Temperature: $\tau_s = 0.1$
- Teacher Momentum: $\tau = 0.996$
- Training Epochs: 400-800

5.5 Superior Performance for Rice Classification

DINO provides the highest expected performance through:

1. **Multi-Scale Learning:** Global crops capture grain shape; local crops capture texture
2. **Transformer Benefits:** Self-attention mechanisms excel at fine-grained feature discovery
3. **Interpretability:** Attention maps reveal discriminative grain regions
4. **State-of-the-Art Results:** Best performance on fine-grained benchmarks

Expected Performance: +5-7% improvement over supervised baseline, representing the highest accuracy among our three methods.

6 Comparative Analysis

Table 1 provides a comprehensive comparison of the three selected methods across key dimensions.

Table 1: Comparative Analysis of Selected SSL Methods

Aspect	**SimCLR**	**BYOL**	**DINO**
Learning Paradigm	Contrastive	Predictive	Distillation
Negative Pairs	Required	Not Required	Not Required
Batch Size	256-512	64-128	64-128
Architecture	CNN	CNN	ViT
Augmentation Sensitivity	High	Medium	Low
Training Stability	Moderate	High	High
Computational Cost	High	Medium	High
Fine-grained Performance	Good	Better	Best
Expected Improvement	+2-4%	+3-5%	+5-7%
Training Time (hours)	8-10	10-12	15-20
GPU Memory (GB)	20-24	12-16	16-20
Interpretability	Low	Low	High

7 Implementation Roadmap

7.1 Phase 1: SimCLR Implementation (Week 1)

- Implement data augmentation pipeline
- Set up ResNet50 encoder with projection head
- Implement NT-Xent loss with large batch sampling
- Train for 200 epochs
- **Expected Outcome:** Baseline SSL model with 2-4% improvement

7.2 Phase 2: BYOL Implementation (Week 2)

- Implement dual-network architecture
- Set up EMA momentum update mechanism
- Implement symmetric prediction loss
- Train for 300 epochs
- **Expected Outcome:** Improved SSL model with 3-5% improvement

7.3 Phase 3: DINO Implementation (Weeks 3-4)

- Implement Vision Transformer architecture
- Set up multi-crop augmentation strategy
- Implement centering and sharpening mechanisms
- Train for 400-600 epochs
- **Expected Outcome:** Best SSL model with 5-7% improvement

8 Expected Outcomes and Impact

Based on literature review and dataset characteristics, we predict the following performance improvements:

Table 2: Performance Predictions

Method	Linear Eval	Fine-Tune	Improvement
Supervised Baseline	-	71-75%	-
SimCLR	68-72%	73-77%	+2-4%
BYOL	71-75%	76-80%	+3-5%
DINO	74-78%	78-82%	+5-7%

8.1 Key Insights Expected

1. **Feature Quality:** SSL methods will learn more robust and generalizable features
2. **Data Efficiency:** Strong performance even with limited labeled data
3. **Fine-grained Understanding:** DINO attention maps will reveal discriminative grain regions
4. **Method Comparison:** DINO will outperform due to multi-scale learning and Transformer architecture

9 Conclusion

This document presents a comprehensive analysis and justification for selecting SimCLR, BYOL, and DINO as our three self-supervised learning methods for rice variety classification. These methods represent different SSL paradigms and provide complementary strengths:

- **SimCLR** establishes a strong contrastive learning baseline
- **BYOL** offers efficient and stable learning without negative pairs
- **DINO** achieves state-of-the-art performance through multi-scale Vision Transformers

The progressive implementation strategy balances computational resources with expected performance gains. The anticipated improvements of 2-7% over our supervised baseline will demonstrate the effectiveness of self-supervised learning for fine-grained agricultural image classification, contributing valuable insights to both the computer vision and agricultural AI communities.

References

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020.

[2] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 21271–21284, 2020.

[3] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

[4] 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," in *International Conference on Learning Representations (ICLR)*, 2021.

[5] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020.

[6] X.-S. Wei, Q. Cui, L. Yang, P. Wang, and L. Liu, "RPC: A large-scale retail product checkout dataset," *arXiv preprint arXiv:1901.07249*, 2021.

[7] A. L. Chandra, S. V. Desai, V. N. Balasubramanian, S. Ninomiya, and W. Guo, "Active learning with point supervision for cost-effective panicle detection in cereal crops," *Plant Methods*, vol. 16, no. 1, pp. 1–16, 2020.

[8] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 702–703, 2020.

[9] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Van Gool, "Scan: Learning to classify images without labels," in *European Conference on Computer Vision (ECCV)*, pp. 268–285, 2020.

[10] L. Zhang, G.-J. Qi, L. Wang, and J. Luo, "Self-supervised visual representation learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4037–4058, 2021.