Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 10
ISET 2025

A comparative study of Contrastive Self-Supervised Learning (CSSL): Methods, Technologies, and Applications

Zhenghan Li*

College of Computer Science and Technology, Hainan Tropical Ocean University, Sanya, Hainan, China

*Corresponding Author: lizhh@stu.hntou.edu.cn

Abstract. With the continuous increase in the cost of data annotation and the explosion of diverse demands, traditional supervised learning is confronted with two major problems: the difficulty in annotating a large number of labels and the expansion bottleneck. Contrastive Self-Supervised Learning (CSSL) provides an effective solution for deep feature extraction in a label-free environment by constructing sample pairs and maximizing their discrimination in the feature space. This review takes Contrastive Predictive Coding (CPC), Simple Framework for Contrastive Learning of Visual Representations (SimCLR), Momentum Contrast (MoCo), Bootstrap Your Own Latent (BYOL), Supervised Contrastive Learning (SupCon), Swapping Assignments between Views (SwAV), and Self-Distillation with No Labels (DINO) as the research objects, systematically sorting out their theoretical frameworks and architectural improvements. The research methods cover InfoNCE loss, momentum encoders, and negative-free self-distillation techniques. This paper focuses on comparing the Top-1 accuracy and computational resource costs of each method in the ImageNet linear evaluation, and presents the core technical differences and performance advantages and disadvantages through three tables. The comparison results show that SupCon and DINO approach or exceed traditional supervised pre-training in different batch settings, while lightweight methods such as BYOL and SwAV perform particularly well in resource-constrained scenarios. Therefore, it can be seen that CSSL has not only made significant progress in feature representation quality and generalization ability, but also laid a solid foundation for subsequent research on the interpretability, multimodal fusion, and efficient deployment of deep learning models.

Keywords: Self-Supervised Learning (SSL); Contrastive learning; Feature representation learning; Model interpretability; Efficient training.

1. Introduction

In the current era of rapid development of artificial intelligence, image recognition, as an important application field of deep learning, has been widely applied in areas such as autonomous driving, medical image analysis, and industrial inspection. However, traditional supervised learning relies on a large amount of manually labeled data. With the increasing complexity of scenarios and the scale of data, the cost of labeling and the risk of error are constantly rising. Currently, there is an urgent need to explore more efficient model learning methods. Self-Supervised Learning (SSL), with its characteristic of not requiring labels and constructing supervisory signals by itself, has rapidly risen in the field of images [1]. Especially Contrastive Self-Supervised Learning (CSSL), by constructing positive and negative sample pairs and maximizing the similarity of positive samples and minimizing the similarity of negative samples in the feature space, has achieved breakthrough progress in semantic understanding and feature extraction [2-4]. Compared with traditional SSL methods, CSSL is more suitable for large-scale unlabeled image sets and has advantages such as strong transferability and flexible structure, making it one of the most valuable research directions at present.

Although there have been continuous innovations in CSSL architecture, loss functions, and training mechanisms, such as Simple Framework for Contrastive Learning of Visual Representations (SimCLR) proposing multi-view data augmentation and Multilayer Perceptron (MLP) projection, Momentum Contrast (MoCo) introducing momentum encoders and negative sample queues, and Bootstrap Your Own Latent (BYOL) eliminating the negative sample mechanism and using

self-distillation structure, there are still many research gaps in CSSL: issues such as insufficient model interpretability, difficulty in adapting to multimodal inputs, and high training resource consumption have not been fully resolved. Therefore, further improving the performance of models in low-resource, complex tasks, and actual deployment scenarios is the key demand of current research [3-5].

This paper will be organized as follows: First, the basic principles and key technologies of contrastive self-supervised learning will be introduced. Then, the architectural differences and performance of representative methods such as Contrastive Predictive Coding (CPC), SimCLR, MoCo, BYOL, Supervised Contrastive Learning (SupCon), Swapping Assignments between Views (SwAV), and Self-Distillation with No Labels (DINO) will be deeply analyzed. The aim is to systematically review the core methods of CSSL from three aspects: theoretical principles, key technologies, and typical experiments, and summarize their advantages and disadvantages through comparative analysis.

2. Theoretical Background

2.1. The foundation of Contrastive Self-Supervised Learning

CPC generates robust and discriminative feature representations by predicting future latent features in the latent space and maximizing the mutual information between samples. The core of this approach lies in the InfoNCE loss function, which distinguishes positive samples from multiple negative samples and enables efficient optimization in a differentiable form [2]. Data-Efficient CPC improves the network architecture and training strategy, including a simplified encoder design and adaptive learning rate scheduling, achieving a Top-1 accuracy of 71.5% on the ImageNet linear evaluation task [6]. DeepCluster employs an alternating clustering and network update mechanism, using clustering results as pseudo-labels to train the model, providing a stable initialization for subsequent contrastive learning methods that rely on clustering structures [7].

2.2. Key Technology Overview

2.2.1. Data Augmentation and View Generation

SimCLR generates multiple different views of the same image by combining various data augmentation strategies, including random cropping, color distortion, and blurring, enabling the model to learn more stable and generalizable features while maintaining view consistency [3].

2.2.2. Momentum Encoder, Memory Queue, and Projection Head Design

MoCo utilizes a momentum-updated encoder and a dynamically maintained queue to treat historical features as a large-scale negative sample pool, thereby overcoming the batch size limitation and enhancing the diversity of negative samples [4]. Both SimCLR and MoCo v2 introduce a nonlinear projection head module that maps intermediate features to the contrastive loss space, reducing the conflict between feature learning and downstream tasks and improving the performance of the final task [3, 8].

2.2.3. Negative Sample-Free Mechanism and Supervised Contrastive Extension

BYOL utilizes mutual learning between an online network and a target network to achieve stable training without explicit negative samples or large-scale memory banks, effectively avoiding the reliance on negative sample sampling in traditional contrastive methods [5]. In the context of labeled data, Supervised Contrastive Learning (SupCon) treats different samples of the same class as positive samples and enhances the aggregation of samples between classes, significantly improving the generalization ability of contrastive learning in supervised training [9].

3. Typical Methods and Method Analysis

3.1. Basic Contrastive Learning Methods (SimCLR and MoCo Series)

SimCLR generates different views of the same image by applying multiple random augmentations and extracts features using a unified ResNet-50 network. After passing through two layers of MLP projections, the features are used to calculate the InfoNCE contrastive loss. With a batch size of 4096 and a long training period, the model achieves a Top-1 accuracy of 76.5% on the ImageNet linear evaluation, comparable to the performance of supervised pre-training of the same scale [3].

MoCo achieves unified comparison between intra-batch samples and historical samples through momentum encoder updates and a dynamic negative sample queue, breaking the batch size limitation on the number of negative samples. Under standard settings, the original MoCo achieves a Top-1 accuracy of 71.1% [4]. MoCo v2 adds a nonlinear MLP projection head and introduces more diverse augmentation strategies, improving performance to 76.6% with a small batch size (256) [8].

3.2. Self-Distillation Contrastive Methods (BYOL and DINO)

BYOL establishes a mutual supervision process between an online network and a target network, updating model parameters through self-distillation without the need for negative samples or memory queues. The online network predicts the projection output of the target network and optimizes it with the mean squared error as the loss function. A base accuracy of 74.3% was achieved on the ResNet-50 architecture, and it could further increase to 79.6% during large-scale training [5].

DINO uses the Vision Transformer (ViT) as its main architecture and conducts unsupervised pre-training through a teacher-student network and a multi-view training process. The teacher network parameters are updated through exponential moving average to ensure the stable convergence of the student network. ViT-Base achieved a Top-1 accuracy of 80.1% in the ImageNet linear test and demonstrated good transferability in downstream tasks such as semantic segmentation [10].

3.3. Supervised and Clustering-based Extension Methods (SupCon and SwAV)

SupCon extends contrastive learning to labeled scenarios by introducing multiple positive sample pairs. For each anchor sample, other samples of the same category are regarded as positive examples, and contrastive loss is used to encourage the clustering of features of the same category. It achieved a Top-1 accuracy of 81.4% in the ImageNet linear evaluation, which is more stable than traditional cross-entropy training and more robust to hyperparameters [9].

SwAV proposed a clustering-based contrastive learning framework that uses online clustering to assign pseudo-labels to different views and makes predictions between views without pairwise comparison. It achieved a Top-1 accuracy of 75.3% in the standard setting and effectively improved the learning of small object features through multi-crop augmentation [11].

4. Analysis and Comparison of Different CSSL Methods

The following Tables 1–3 compare the core technologies, performance on ImageNet, advantages and disadvantages, and applicable scenarios of different CSSL methods.

Table 1. Core Techniques of CSSL Methods

Method	Core Techniques	Negative Sample Requirement	Additional Components
CPC	Predictive InfoNCE	In-batch sampling	None
SimCLR	Enhanced data augmentation + nonlinear MLP projection head	Large-batch contrastive learning	None

Continue Table 1

MoCo	Momentum encoder + dynamic negative sample queue	Requires memory bank	Momentum encoder
MoCo v2	MoCo + nonlinear projection head + improved augmentation	Requires memory bank	Momentum encoder, MLP projection head
BYOL	Online–target self-distillation, no negative samples	None	Target network
SupCon	Supervised contrastive with multiple positive samples	Same-class positive pairs	Label information
SwAV	Clustering-based view-swapping prediction	No explicit negatives required	Cluster prototypes
DINO	Teacher–student self-distillation + ViT architecture	No negative samples	Vision Transformer

Table 2. ImageNet Linear Evaluation Performance Comparison (ResNet-50 / ViT-Base)

Method	Top-1 Accuracy	Training Batch Size
CPC	48.7% → 71.5%	256
SimCLR	76.5%	4096
MoCo	71.1%	256 (queue length: 8192)
MoCo v2	76.6%	256
BYOL	74.3% → 79.6%	4096
SupCon	81.4%	256
SwAV	75.3%	256
DINO	80.1%	512

Table 3. Advantages, Limitations, and Applicable Scenarios

Method	Advantages	Limitations	Applicable Scenarios
CPC	Directly maximizes mutual information; scalable to multimodal tasks	Requires well-designed prediction tasks	Cross-modal learning (speech, text, image)
SimCLR	Simple to implement; easy to reproduce	Highly dependent on large batch sizes; resource-heavy	Abundant unlabeled data with sufficient resources
MoCo	High diversity of negatives; more stable training	Requires fine-tuning momentum and queue size	Contrastive learning with small/medium batches
MoCo v2	Better convergence under small batch sizes	Still relies on momentum encoder and memory queue	Performance boost in small/medium batch environments
BYOL	No negative sampling; low memory overhead	Unstable training; requires long training and tuning	Resource-constrained; exploring self-distillation

Method	Advantages	Limitations	Applicable Scenarios
SupCon	Enhanced supervision signal; high accuracy	Requires labels; semi-supervised setting needs design	Few-shot with labels; robust to hyperparameters
SwAV	Novel clustering-based contrast; memory & compute friendly	Risk of cluster collapse; sensitive to hyperparameters	Mid-scale unsupervised learning under memory limits
DINO	Works well with Transformers; strong downstream transfer	ViT is compute-intensive and memory-demanding	Transfer tasks like segmentation, detection

5. Discussion and Future Prospects

5.1. Model Interpretability and Trustworthiness

Although contrastive self-supervised learning methods excel in unlabeled feature extraction, their decision-making processes are generally characterized as "black boxes," which hinders the assessment of the trustworthiness of model outputs and risk control. In the future, visualization techniques (such as Grad-CAM, attention maps) and feature contribution analysis from a game theory perspective (such as Shapley values) can be combined to reveal the specific influence mechanisms of different augmented views or samples on training objectives. Additionally, developing a contrastive learning framework based on causal inference can help construct feature representations that are more in line with human cognitive logic, thereby enhancing the application safety of models in critical fields such as healthcare and autonomous driving.

5.2. Multimodal Fusion and Adaptation to Downstream Tasks

Most current CSSL methods focus on a single image modality, but real-world applications often involve multiple data sources such as images, text, and audio. Future research should explore cross-modal contrastive objective functions to map different modality views into a unified feature space, enhancing semantic consistency. For instance, the Contrastive Captioner (CoCa) model achieves multimodal alignment between images and text through a joint contrastive and generative loss, significantly improving performance in tasks such as image classification, retrieval, and question answering [12]. Moreover, studies have shown that when transferring to downstream tasks such as object detection, instance segmentation, and 3D reconstruction, designing lightweight fine-tuning strategies (such as freezing some parameters and introducing small-scale adaptation modules) can enhance model adaptability while maintaining speed and accuracy.

5.3. Efficient Training and Resource-Constrained Deployment

Although lightweight designs like BYOL and SwAV have shown potential on edge devices, contrastive learning still heavily relies on large-scale training and long-term optimization. Future efforts can be made in two directions: First, introducing adaptive negative sample sampling and online clustering mechanisms to reduce computational redundancy. For example, the Adaptive Sampling (AdaS) method improves negative sample selection in contrastive learning by evaluating their importance, enhancing model performance [13]. Second, combining neural network compression techniques (quantization, pruning, knowledge distillation) to reduce computational and storage costs during inference while maintaining core contrastive signals, promoting the large-scale application of CSSL models in mobile and industrial embedded systems.

6. Conclusion

Contrastive Self-Supervised Learning has become an important research direction in the field of image recognition. From CPC’s InfoNCE to SimCLR’s augmentation strategies, MoCo’s momentum

queue, BYOL's negative-free mechanism, and SwAV and DINO's clustering and Transformer innovations, a series of methods have continuously pushed the boundaries of supervised pre-training. In the future, with the deepening of interpretability research, it may be able to reveal the semantic information contained in self-supervised representations. Multimodal fusion techniques (such as visual-language contrastive learning) will endow models with stronger cross-domain understanding capabilities. Efficient training algorithms (such as more lightweight contrastive losses and improved regularization strategies) are expected to reduce computational costs. Combining federated learning and privacy protection mechanisms can enable CSSL to be applied in data-sensitive fields such as healthcare and security. It is believed that under the impetus of these directions, self-supervised visual pre-training will truly enter a new era of ubiquitous intelligence without the need for large-scale labeled data, bringing broader applications and deeper impacts to computer vision.

References

[1] L. Jing, Y. Tian, Self-supervised visual feature learning with deep neural networks: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2020) 4037–4058.

[2] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).

[3] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: Proc. Int. Conf. Mach. Learn. (ICML), PMLR, 2020, pp. 1597–1607.

[4] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 9729–9738.

[5] J.B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, et al., Bootstrap your own latent: a new approach to self-supervised learning, Adv. Neural Inf. Process. Syst. 33 (2020) 21271–21284.

[6] O. Henaff, Data-efficient image recognition with contrastive predictive coding, in: Proc. Int. Conf. Mach. Learn. (ICML), PMLR, 2020, pp. 4182–4192.

[7] M. Caron, P. Bojanowski, A. Joulin, M. Douze, Deep clustering for unsupervised learning of visual features, in: Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 132–149.

[8] X. Chen, H. Fan, R. Girshick, K. He, Improved baselines with momentum contrastive learning, arXiv preprint arXiv:2003.04297 (2020).

[9] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, et al., Supervised contrastive learning, Adv. Neural Inf. Process. Syst. 33 (2020) 18661–18673.

[10] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2021, pp. 9650–9660.

[11] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, A. Joulin, Unsupervised learning of visual features by contrasting cluster assignments, Adv. Neural Inf. Process. Syst. 33 (2020) 9912–9924.

[12] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, Y. Wu, CoCa: contrastive captioners are image-text foundation models, arXiv preprint arXiv:2205.01917 (2022).

[13] S. Wan, Y. Zhan, S. Chen, S. Pan, J. Yang, D. Tao, C. Gong, Boosting graph contrastive learning via adaptive sampling, IEEE Trans. Neural Netw. Learn. Syst. (2023).