International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 10 Issue 10, October - 2026, Pages: 221-225

Self-Supervised Learning in Computer Science: A Review of Contrastive, Non-Contrastive and Masked Modeling Approaches

Dr. Virendra Tank

Shri Mahaveer College, Jaipur, Rajasthan, India

vtank87@gmail.com 0009-0003-9126-0982

Abstract*— *Deep learning has traditionally relied on large, manually labeled datasets, which are expensive to build and difficult to scale. Self-supervised learning (SSL) removes this dependence by creating supervisory signals from the structure of the data itself. This paper reviews the main SSL families used in computer science: pretext-task methods, contrastive learning, non-contrastive (self-distillation and redundancy-reduction) methods, clustering-based methods, and masked modeling. We describe the methodology of each family, identify the central problems (negative-sample cost, representation collapse, augmentation dependence and scalability), discuss the solutions proposed in the literature, and compare reported results on ImageNet. Applications in vision, language, speech, source-code understanding and educational data mining are also discussed. We conclude with open challenges and future directions, including joint-embedding predictive architectures, multimodal SSL and data-efficient foundation models.

Index Terms— self-supervised learning, contrastive learning, masked autoencoders, representation learning, foundation models, deep learning.

I. INTRODUCTION

Supervised deep learning has produced remarkable results in vision, language and speech [1], but its progress is closely tied to large annotated corpora such as ImageNet [2]. Annotation is slow, costly and often requires domain experts, while unlabeled data (images, text, audio, source code, student interaction logs) is available in enormous quantities. Self-supervised learning addresses this gap: a model is first pre-trained on a *pretext* objective whose labels are generated automatically from the data, and the learned representation is later transferred to downstream tasks using little or no labeled data [3], [4].

The idea became dominant in natural language processing with BERT [19] and GPT-style models [26], and in the last few years it has matched or surpassed supervised pre-training in computer vision [10], [11], [22]. SSL is now the base technique behind foundation models [29]. This review is aimed at students and researchers who need a structured overview of the field. The main contributions are: (i) a taxonomy of SSL methods; (ii) an analysis of the key problems and the solutions that addressed them; (iii) a consolidated comparison of reported results; and (iv) a discussion of applications in computer science and education and of future directions.

The rest of the paper is organized as follows. Section II describes the review methodology and the SSL families. Section III discusses the main problems and Section IV the corresponding solutions. Section V compares results, Section VI covers applications, and Sections VII and VIII give the conclusion and future directions.

II. METHODOLOGY

A. Review Methodology

Papers were selected from major venues (NeurIPS, ICML, ICLR, CVPR, ICCV, ACL/NAACL, IEEE TPAMI and TKDE) and from surveys [3]–[5]. Priority was given to works that introduced a new SSL principle or that substantially changed benchmark results. Methods are grouped by the type of learning signal they use, and results are taken from the original papers, mainly the ImageNet-1k linear-evaluation and fine-tuning protocols.

B. Pretext-Task Methods

Early SSL designed hand-crafted tasks whose solution requires semantic understanding: predicting the relative position of image patches [6], solving jigsaw puzzles [8], and predicting image rotation [7]. Their weakness is that the learned features are specific to the task, which limits transfer.

C. Contrastive Learning

Contrastive methods pull together representations of two augmented views of the same sample (positives) and push apart views of different samples (negatives). Contrastive Predictive Coding [9] introduced the InfoNCE loss, which for an anchor z with positive z^+ and N negatives is $L = -\log(\exp(\mathrm{sim}(z, z^+)/\tau) / \Sigma \exp(\mathrm{sim}(z, z_i)/\tau))$, where τ is a temperature. SimCLR [11] showed that strong augmentation, a non-linear projection head and large batches are critical. MoCo [10], [12] replaced large batches with a queue of negatives encoded by a momentum encoder, making training feasible on modest hardware.

D. Non-Contrastive Methods

BYOL [13] showed that negatives are not required: an online network predicts the output of a slowly updated target network. SimSiam [14] achieved similar results without a momentum encoder, using only a stop-gradient. Barlow Twins [15] and VICReg [16] avoid collapse through explicit regularization, by decorrelating feature dimensions or by enforcing variance and covariance constraints. DINO [18] applies self-distillation to vision transformers and produces features whose attention maps segment objects without supervision.

E. Clustering-Based Methods

SwAV [17] compares views through online cluster assignments instead of pairwise comparisons, which reduces memory cost and allows multi-crop training with views of different sizes.

F. Masked Modeling

Inspired by BERT [19], masked image modeling hides part of the input and trains the model to reconstruct it. BEiT [21] predicts discrete visual tokens, while the Masked Autoencoder (MAE) [22] masks 75% of image patches, encodes only the visible ones and reconstructs pixels with a light decoder. This asymmetric design makes pre-training roughly three times faster or more and scales well to large Vision Transformers [20]. I-JEPA [24] predicts representations of masked regions rather than pixels, avoiding the need for hand-designed augmentations.

TABLE I. Taxonomy of self-supervised learning families

Family	Representative methods	Core idea	Negatives needed?
Pretext tasks	Context prediction [6], Jigsaw [8], Rotation [7]	Solve a hand-designed task with automatic labels	No
Contrastive	CPC [9], MoCo [10], SimCLR [11]	Attract positive pairs, repel negatives (InfoNCE)	Yes
Self-distillation	BYOL [13], SimSiam [14], DINO [18]	Predict a target network output; stop-gradient / momentum avoids collapse	No
Redundancy reduction	Barlow Twins [15], VICReg [16]	Decorrelate features; regularize variance and covariance	No
Clustering	SwAV [17]	Match online cluster assignments across views	No
Masked modeling	BEiT [21], MAE [22], I-JEPA [24]	Reconstruct or predict masked parts of the input	No

III. PROBLEMS AND CHALLENGES

Cost of negatives. Contrastive objectives benefit from many negative samples, which requires very large batches or memory banks [10], [11]. This raises computational cost and limits access for smaller labs.

Representation collapse. Without negatives, a network can trivially output a constant vector for every input. Collapse can be complete or partial (dimensional collapse), where embeddings occupy only a low-dimensional subspace [31].

Dependence on augmentations. Invariance-based methods learn whatever invariances the augmentations encode; poorly chosen views remove useful information, and good views are often domain-specific [32]. This is a particular difficulty in domains such as text, tabular data or medical images, where standard image augmentations do not apply.

Scalability and evaluation. Pre-training large models is expensive, and linear probing may underestimate the quality of features learned by masked modeling, which often needs fine-tuning [22]. Different papers also use different backbones, epochs and protocols, which makes fair comparison difficult [5].

Bias and robustness. Models pre-trained on uncurated web-scale data can inherit social biases and may be fragile under distribution shift [29].

IV. SOLUTIONS

Reducing the need for large batches. MoCo maintains a dynamic dictionary using a momentum encoder [10], and MoCo v2 adopts SimCLR-style augmentation and a projection head to reach strong accuracy with ordinary batch sizes [12]. SwAV avoids pairwise comparison altogether [17].

Avoiding collapse. BYOL uses a predictor and an exponential-moving-average target [13]; SimSiam shows that a stop-gradient alone is sufficient [14]. Barlow Twins pushes the cross-correlation matrix of two views towards identity [15], and VICReg adds explicit variance and covariance terms [16]. Wang and Isola explain contrastive learning through two properties, *alignment* of positives and *uniformity* of the embedding distribution on the hypersphere [30], which gives a principled view of why collapse is prevented.

Better views and weaker augmentation dependence. Studies on view selection show that the best views share task-relevant information while discarding nuisance factors [32]. Masked modeling [21], [22] and predictive architectures such as I-JEPA [24] reduce reliance on hand-crafted augmentations, because the learning signal comes from the input structure itself.

Efficient and scalable pre-training. MAE processes only about a quarter of the patches in the encoder, giving large speed-ups and allowing ViT-Huge to be trained on ImageNet-1k alone [22]. DINOv2 combines self-distillation with masked objectives on a curated dataset to produce general-purpose frozen features [23]. Large pre-trained SSL models also act as strong semi-supervised learners when only a small portion of labels is available [11].

V. RESULTS AND COMPARATIVE ANALYSIS

Table II summarizes ImageNet-1k top-1 accuracy as reported in the original papers. Linear evaluation freezes the encoder and trains a linear classifier; fine-tuning updates all weights. Because training budgets, augmentations and epochs differ between papers, the numbers show trends rather than strict rankings.

TABLE II. Reported ImageNet-1k top-1 accuracy of representative methods

Method	Backbone	Paradigm	Top-1 (%)	Protocol / Source
Supervised	ResNet-50	Labels	76.5	Supervised baseline reported in [13]
SimCLR	ResNet-50	Contrastive	69.3	Linear eval. [11]
MoCo v2	ResNet-50	Contrastive	71.1	Linear eval. [12]
SimSiam	ResNet-50	Self-distillation	71.3	Linear eval. [14]
Barlow Twins	ResNet-50	Redundancy red.	73.2	Linear eval. [15]
VICReg	ResNet-50	Regularization	73.2	Linear eval. [16]
BYOL	ResNet-50	Self-distillation	74.3	Linear eval. [13]
SwAV	ResNet-50	Clustering	75.3	Linear eval. [17]
MoCo v3	ViT-B	Contrastive	76.7	Linear eval. [28]
DINO	ViT-B/8	Self-distillation	80.1	Linear eval. [18]
DINOv2	ViT-g/14	Distillation + masking	86.5	Linear eval. [23]
MAE	ViT-L/16	Masked modeling	85.9	Fine-tuning [22]
MAE	ViT-H/14	Masked modeling	86.9	Fine-tuning [22]

Note: values are taken from the cited papers and use different training schedules; fine-tuning and linear-evaluation rows should not be compared directly.

Three observations follow. First, on ResNet-50 the gap between SSL and supervised training has narrowed from about 25 points for early contrastive work to about one to seven points, with SwAV and BYOL close to the supervised baseline [13], [17]. Second, non-contrastive methods match or exceed contrastive ones without negatives, which shows that negatives are not essential. Third, with Vision Transformers, masked modeling and self-distillation scale better than earlier contrastive designs, and the strongest reported results use large ViTs [22], [23]. SSL also gives clear benefits in low-label regimes and in transfer to detection, segmentation and other downstream tasks [10], [22].

VI. APPLICATIONS IN COMPUTER SCIENCE AND EDUCATION

Vision and multimodal learning. SSL features are used for classification, detection, segmentation and retrieval [10], [22], and language–image pre-training such as CLIP uses a contrastive objective between images and captions [27]. **Natural language and speech.** BERT [19] and GPT-3 [26] rely on masked or next-token prediction, while wav2vec 2.0 learns speech representations by contrastive prediction over masked latent frames [25]. **Source code.** CodeBERT [33] pre-trains on code and natural language, and

models trained on code support automatic programming assistance [35]; these tools are increasingly used in programming education for feedback and hint generation.

Educational data mining. In educational settings labeled outcomes are scarce, whereas logs of student interactions are abundant. Contrastive learning has been used for knowledge tracing, where augmented interaction sequences are used to learn more robust student representations and to improve prediction of future performance [34]. Similar ideas apply to dropout prediction and personalized recommendation of learning material, where SSL can reduce the need for labeled outcomes.

VII. CONCLUSION

Self-supervised learning has moved from a niche idea to the main way of building representations for modern deep learning. Contrastive methods demonstrated that label-free pre-training can rival supervised training; non-contrastive and clustering methods removed the dependence on negatives; and masked modeling made SSL efficient and scalable for Vision Transformers. Reported results show that SSL-pretrained models now approach or exceed supervised baselines on ImageNet while needing far fewer labels for downstream tasks. Problems such as collapse, augmentation dependence, expensive pre-training and non-uniform evaluation have been partly solved but are not fully resolved. SSL is also beginning to benefit programming and education research, where unlabeled data is plentiful and labels are scarce.

VIII. FUTURE DIRECTIONS

1) Predictive and world-model architectures. Joint-embedding predictive methods such as I-JEPA [24] learn in representation space rather than pixel space, and extending them to video and sequential data is a promising direction. **2) Multimodal and unified SSL.** Combining text, image, audio and code in one pre-training objective can give more general representations [27], [29]. **3) Theory.** A better theoretical understanding of why SSL works, including what is learned by invariance versus reconstruction, is still limited [5], [30]. **4) Efficiency and accessibility.** Smaller models, data-efficient pre-training and distillation are needed so that SSL is usable outside large industry labs. **5) Standardized evaluation.** Common backbones, budgets and benchmarks would make comparisons fair [5]. **6) Responsible SSL.** Auditing bias, privacy and robustness of models trained on web-scale data is essential [29]. **7) Education.** Privacy-preserving SSL on learner data, SSL for programming-feedback systems and cross-course transfer in knowledge tracing remain open problems.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE CVPR*, 2009, pp. 248–255.

[3] X. Liu *et al.*, "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, 2023.

[4] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, 2021.

[5] R. Balestriero *et al.*, "A cookbook of self-supervised learning," *arXiv:2304.12210*, 2023.

[6] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE ICCV*, 2015, pp. 1422–1430.

[7] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. ICLR*, 2018.

[8] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. ECCV*, 2016, pp. 69–84.

[9] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv:1807.03748*, 2018.

[10] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF CVPR*, 2020, pp. 9729–9738.

[11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, PMLR 119, 2020, pp. 1597–1607.

[12] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv:2003.04297*, 2020.

[13] J.-B. Grill *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," in *Proc. NeurIPS*, vol. 33, 2020, pp. 21271–21284.

[14] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF CVPR*, 2021, pp. 15750–15758.

[15] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow Twins: Self-supervised learning via redundancy reduction," in *Proc. ICML*, PMLR 139, 2021, pp. 12310–12320.

[16] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," in *Proc. ICLR*, 2022.

[17] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. NeurIPS*, vol. 33, 2020, pp. 9912–9924.

[18] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF ICCV*, 2021, pp. 9650–9660.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[20] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.

[21] H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," in *Proc. ICLR*, 2022.

[22] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF CVPR*, 2022, pp. 16000–16009.

[23] M. Oquab *et al.*, "DINOv2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res.*, 2024.

[24] M. Assran *et al.*, "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proc. IEEE/CVF CVPR*, 2023, pp. 15619–15629.

[25] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, vol. 33, 2020, pp. 12449–12460.

[26] T. Brown *et al.*, "Language models are few-shot learners," in *Proc. NeurIPS*, vol. 33, 2020, pp. 1877–1901.

[27] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, PMLR 139, 2021, pp. 8748–8763.

[28] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. IEEE/CVF ICCV*, 2021, pp. 9640–9649.

[29] R. Bommasani *et al.*, "On the opportunities and risks of foundation models," *arXiv:2108.07258*, 2021.

[30] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *Proc. ICML*, PMLR 119, 2020, pp. 9929–9939.

[31] L. Jing, P. Vincent, Y. LeCun, and Y. Tian, "Understanding dimensional collapse in contrastive self-supervised learning," in *Proc. ICLR*, 2022.

[32] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" in *Proc. NeurIPS*, vol. 33, 2020, pp. 6827–6839.

[33] Z. Feng *et al.*, "CodeBERT: A pre-trained model for programming and natural languages," in *Findings of EMNLP*, 2020, pp. 1536–1547.

[34] W. Lee, J. Chun, Y. Lee, K. Park, and S. Park, "Contrastive learning for knowledge tracing," in *Proc. ACM Web Conf. (WWW)*, 2022, pp. 2330–2338.

[35] M. Chen *et al.*, "Evaluating large language models trained on code," *arXiv:2107.03374*, 2021.