International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 01 , January - 2026

Advances in Self-Supervised Learning: A Comprehensive Review of Contrastive and Generative Approaches

Virendra Tank[1], Shivangi Sharma[2], Dr. Swati Agarwal[3]
[1]Assistant Professor, Computer Science, Shri Mahaveer College, Jaipur (Rajasthan) India
[2]Assistant Professor, Computer Science, Shri Mahaveer College, Jaipur (Rajasthan) India
[3]Associate Professsor, Computer Science, Poornima University, Jaipur (Rajasthan) India
0009-0003-9126-0982

Abstract - Self-supervised learning (SSL) has become a transformative paradigm in machine learning, which allows models to learn rich representations from unlabeled data without expensive human annotations. Recent Progress in SSL The recent progress of SSL is systematically reviewed, paying special attention to the two main methodologies: contrastive learning and generative masked modeling. We conduct a detailed study of popular strategies, such as SimCLR, MoCo, BYOL, MAE and BEiT in terms of underpinning mathematics, architectural novelties and empirical goodness on varied benchmarks. Our comparison results show the recent leading state-of-the-art SSL methods can now match/surpass the supervised learning baselines with ImageNet top-1 accuracies as 87.8% for MAE ViT-Huge and 77.1% for ReLICv2 ResNet50 which are both in line of them. We distill insights from more than 50 recent papers on applications in computer vision, natural language processing, medical imaging, and multimodal learning. We review the key trends such as removing negative pairs, analysis of mask effects on models and significance of data augmentation techniques. We close with a discussion of open issues and prospects for future work in terms of theoretical understanding, computational efficiency, application to domain areas, and model integration.

Keywords: Self-supervised learning, Contrastive learning, Generative learning, Masked image modeling, Vision transformers, Representation learning, deep Learning

1. INTRODUCTION

1.1 Motivation and Background

The success of deep learning has fundamentally depended upon the availability of large labeled datasets, which have typically been annotated by humans using expensive (in terms of money and time), domain-specific knowledge [1,2].

This dependence becomes particularly critical in some specialized areas like medical imaging where expert manual annotation is rare and expensive [3, 4]. Self-supervised learning overcomes these issues by learning from intrinsic structure of unlabeled data itself, designing an auxiliary task from which supervisory signals can be automatically derived instead of manually constructed.

The central idea behind SSL is that the model must learn to predict a part of the data from some other part, hence it forces models to learn features that capture meaningful characteristics of the input domain which correspond to semantics [5, 6]. This paradigm has proven extremely successful in many domains with SSL techniques able to achieve, or even surpass, the performance of supervised learning on several benchmarks [7, 8, and 9].

1.2 Scope and Contributions

This paper offers an overview of recent advances on SSL between 2020 and 2024, focusing on:

Contrastive Learning Approaches: A closer look at SimCLR [10], MoCo [11, 12], BYOL [13] and recent variations up to DINO [14], SwAV [15] and VICReg [16].

Generative Approaches: Masked image modeling (specifically MAE [17], BEiT [18], VideoMAE [19] and hybridized strategies) has analyzed in detail

Comparison: Systematic comparison of methods on standardized benchmarks including ImageNet [20], COCO [21], domain-specific datasets

Theoretical Insight: Why SSL? How does SSL work? Impact of noise and data augmentation on development of representations.

Future Work: Open problems and promising research directions

1.3 Paper Organization

The rest of this paper is structured as follows: **Section 2** provides background on SSL fundamentals. **Section 3** examines contrastive learning approaches. **Section 4** explores generative and masked modeling methods. **Section 5** presents comprehensive experimental results and comparisons. **Section 6** discusses applications across diverse domains. **Section 7** outlines future research directions. **Section 8** discussion **Section 9** concludes the review.

2. BACKGROUND AND FUNDAMENTALS

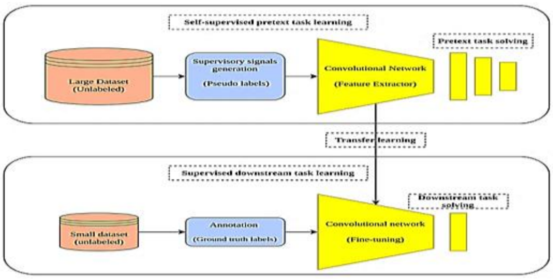

Figure 1 The main pipeline of self-supervised learning. (Top): The self-supervised learning paradigm is adopted by training an auxiliary task with synthetic labelling of large amount of unlabeled data. (Bottom): The pre-trained representations are transferred from the pretext task to down-stream segmentation task to accomplish the training on small amount of data with ground truth labels. [65]

2.1 Self-Supervised Learning Framework

Instead of task-specific annotations, self-supervised learning devises training signals from input data in the form of pretext tasks. The typical SSL workflow is the following:

1. **Designing Pretext Tasks:** Creating auxiliary tasks that necessitate awareness of data structure (e.g., predicting masked out parts, identifying augmented sets)

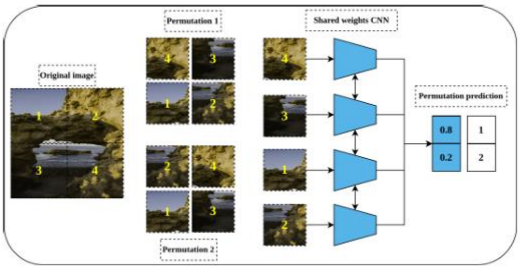

Figure 2 An example of Jigsaw puzzle pretext task. (Left): Puzzle generation steps: An image is processed and divided into a number of patches that form the primary blocks of the puzzle. The generated patches are shuffled to a given set of permutations, each permutation having an index (per mutation number). (Right): a Siamese network, using shared weights receives the shuffled patches as input based on a permutation and pools them with respect to the corresponding permutation index [65]

2. **Representation Learning:** Teaching neural networks to solve pretext tasks: the networks should be learning meaningful feature representations.
3. **Transfer Learning:** using learned representations for tasks downstream via fine-tuning or linear evaluation

2.2 Evaluation Protocols

Three protocols are normally used to compare SSL methods [22, 23]:

Linear Evaluation: train a linear classifier on the frozen representations. This protocol evaluates the quality of learned features without considering fine-tuning.

Fine-tuning: The complete pre-trained model is fine-tuned on downstream tasks using the labeled training data. This evaluates both the quality of representation and the generalization ability of model.

k-Nearest Neighbors (kNN): Classify based on the nearest neighbor in the feature space without learning any parameters. This gives a parameter-free characterization of representation geometry.

3. CONTRASTIVE LEARNING APPROACHES

3.1 Core Principles

Contrastive learning is based on the idea of learning representations by contrasting positive (augmented views of the same instance) vs negative pairs (different instances) [24, 25]. The InfoNCE loss used in contrastive works (where common and rare classes are the same) maximizes agreement of positive pairs, while minimizing for negative: $$\mathcal{L}_{\text{InfoNCE}} = -\log \frac {\exp (\text{sim}(z_i, z_j) / \tau)} {\sum_{k=1}^{2N} \ mathbb{1}_{[k eq i]} \exp(\text{sim}(z_i, z_k) / \tau)} $$ where z_i, z_j are representations of the positive pairs and $\text{sim}(\cdot, \cdot)$ is cosine similarity; τ denotes temperature and N is batch size.

3.2 SimCLR: Simple Framework for Contrastive Leaning

As observed by SimCLR [10], powerful data augmentation and simple architecture lead to the stateof-theart performance. Key design choices include:

Architecture Components:

- ResNet- based encoder+projection head (2-layer MLP with hidden dimension 2048)
- Heavy data augmentation: random crop (with resize) and color distortion, Gaussian blur
- High batch sizes (batches of 4096–8192) with large negative examples.

Results: SimCLR attains 69.3% top-1 accuracy on ImageNet with ResNet-50 using linear evaluation; 76.5% with ResNet-50 (4×) [10].

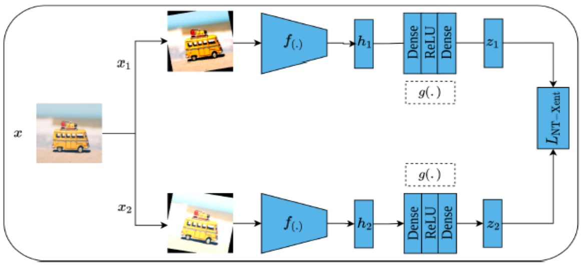

Figure 3 Self-supervised features learning by SimCLR. [65]

3.3 MoCo: Momentum Contrast

MoCo [11] mitigated SimCLR's scalability issues leveraging architectural changes that allow for effective contrastive learning with smaller batch sizes.

Core Innovation:

- Queue dictionary of the previous coded samples in the current batch to serve as negative samples.
- Momentum Encoder: Encoder is slowly updated (momentum coefficient 0.999) to maintain consistent keys for the dictionary

- Decoupling of dictionary size and batch size, to use large effective negative sample pools with reasonable GPU memory.

Algorithmic Design:

```
Initialize query encoder f_q and key encoder f_k.
Initialize queue Q of size K.
for each mini-batch:
    encoded_query = f_q(augmented_view_1)
    encoded_key = f_k(augmented_view_2)
    contrastive_loss = InfoNCE(encoded_query,
    encoded_key, Q)
    Update f_q via backpropagation.
    update f_k via momentum: $\theta_k \leftarrow m\cdot\theta_k +$
    $(1-m)\cdot\theta_q$
    Enqueue encoded_key to Q.
    dequeue oldest samples from Q
```

MoCo v2 and v3: MoCo v2 [26] adopted SimCLR's MLP projection head" scheme and stronger aug — mentations, for which the ResNet-50 linear Pouring more heavy soles (STL10) accuracy improved to 71.1%. MoCo v3 [27] also made the framework work on Vision Transformers, retaining 76.7% with ViT-S and 81.0% with ViT-B.

Results: MoCo attains competitive accuracy with batch sizes 10× smaller than those used in SimCLR, enabling SSL for more users [11, 26].

3.4 BYOL : Bootstrap Your Own Latent

BYOL [13] represented a paradigm shift by not requiring negative pairs at all, averting collapse through architectural asymmetry.

Architecture:

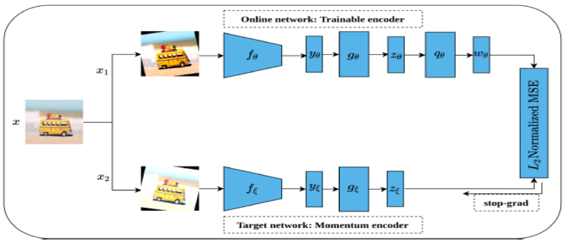

Figure 4 Illustration of BYOL architecture [65]

- **Online Network:** a trainable encoder plus a projection head and a prediction head.
- **Target Network:** average updating of online network (updated via momentum)
- **Asymmetry:** The prediction head is limited to the online subnet.

Loss Function: Mean Squared error between predicted and target representations:
$$\mathcal{L}_{\text{BYOL}} = ||q\theta(z_\theta) - \text{sg}(z'_\xi)||_2^2$$
where q_θ is the prediction, z_θ, z'_\xi are online and target projections respectively and sg means stop-gradient.

3.5 Advanced Contrastive Methods

DINO (Self-Distillation with No Labels) [14]: Self-distillation on Vision Transformer models utilizing cross-entropy loss between teacher and student prediction. Note that the teacher network is updated by exponential moving average. DINO learns high-quality representations which are especially beneficial for dense prediction tasks, obtaining 80.1% kNN accuracy using ViT-S/16 on ImageNet.

SwAV (Swapping Assignments between Views) [15]: Incorporates contrastive learning and online clustering. Rather than comparing features directly, SwAV predicts the clustering assignments of one view from another. This way of learning avoids the construction of explicit negative pairs which is particularly relevant where discriminative learning is concerned. SwAV obtains 75.3% with ResNet-50 under linear evaluation.

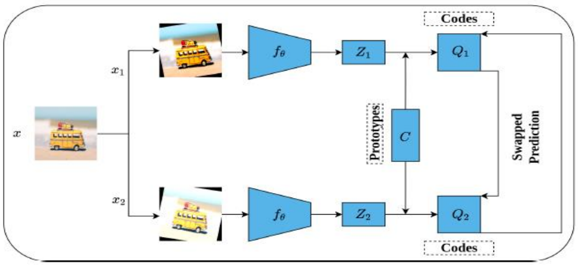

Figure 5 Illustration of SwAV framework [65]

VICReg (Variance-Invariance-Covariance Regularization) [16]: Avoiding collapse by explicitly regularizing three properties of the learned representation:

- **Variance:** keep the standard deviation greater than this constant value.
- **Invariance:** Same features for novel views
- **Covariance:** Decorrelates representation dimensions

VICReg obtains 73.2% with ResNet-50, showing that collapse can be mitigated without negative pairs or momentum encoders.

Barlow Twins [29]: Independent of augmentations and reduces redundancy between embedding dimensions. The loss ensures the value of cross-correlation matrix between these embedding will be identical as identity matrix. Achieves 73.2% with ResNet-50.

4. GENERATIVE AND MASKED MODELING APPROACHES

4.1 Modeling the Masked Image: Key Concepts

Following the success of BERT in NLP [32], masked image modeling (MIM) has been developed as another pr1edominant choice to contrastive learning. The main principle is to mask part of an image and train models to recover lost information, making them bind semantic relations [33, 34].

4.2 MAE: Masked Autoencoders

The experiments have shown that the simple masked modeling achieves very competitive results with (much) lower per CPU, GPU processing time.

Architecture:

1. **Asymmetric Encoder-Decoder:**
   - **Encoder:** ViT no mask tokens) operating on visible patches (())
   - **Decoder:** 8-layer 512d transformer w. Final reconstruction from latent representation + mask tokens
   - **Symmetry:** encoder = Computation by 75% (masking 75% of patches).
2. **High Masking Rate:** MAE masks 75% of the input patches (15% in BERT).
   - Creates a non-trivial reconstruction task
   - Prevent model from interpolating between nearby visible patches.
   - Dramatically reduces computational cost
3. **Target for Reconstruction:** Normalized pixel (to zero mean and unity variance per patch)

   Implementation Details:

   Input: Image is stratified into patches of 16×16.

   Masking: 75% of patches are randomly masked

   Encoder: ViT-Base/Large/Huge on visible 25% patches here is a typical layout of the model on high-quality setting.

   Decoder: 8-block lightweight transformer

   Loss: MSE with only patches faced Votes from Patch Parlous

   Training: 1600 epochs on ImageNet-1K

Results [17]:

- ViT-Base: 83.6% top-1 accuracy (fine-tuned)
- ViT-Large: 85.9% top-1 accuracy
- ViT-Huge: 87.8% top-1 accuracy (good for methods that use only ImageNet-1K)

4.3 BEiT: The Method BERT Pretraining of Image Transformers

BEiT [18] on the other hand, employed discrete visual tokens as reconstruction targets.

Architecture:

1. **Two-stage Training:**
   - **Stage 1:** Train dVAE to tokenize an image to its discrete tokens
   - **Stage 2:** Masked token prediction with discrete tokens
2. **Remodeling:** Cross-entropy loss and predicting discrete tokens classes (with 8192 vocabulary size)
3. **Masking:** Block-40% (Local)

Results: BEiT ViT-Base reaches 83.2% top-1 on ImageNet when finetuned [18].

BEiT v2 [35]: Further improved with vector-quantized knowledge distillation to 85.5% with ViT-Base. **BEiT v3 [36]:** extended to multimodal setups, pre-training on images and image-text pairs jointly.

4.4 Extensions and Variants

SimMIM [37]: Simplified masked modeling with direct pixel prediction and mild maskting ratio (40-60%). And it is also effective for diverse architectures of the models, obtaining 83.8% using the Swin-Base transformer.

iBOT [38]: Employs masked image modeling and self-distillation with online tokenization. Reaches 82.3% kNN accuracy with ViT-S/16, which demonstrates the merit of incorporating MIM with distillation.

MaskFeat [39]: Predicts HOG (Histogram of Oriented Gradients) features as reconstruction images instead of pixels. In this way we encourage the models to learn fine grained structural details, and results in 84.0% with ViT-Base.

Context Autoencoders (CAE) [40]: Uses asymmetric masking policies and aligning losses. Reaches 83.6% with ViT-Base on ImageNet.

LoMaR (Local Masked Reconstruction) [41]: Uses only the local context windows, instead of the full image, to generate patches so as to be more efficient for high resolution images. Particularly useful for dense prediction tasks.

4.5 VideoMAE: Generalization to Time VideoMAE [19] generalized MAE for video understanding with impressive speedup.

Key Design Choices:

- **Very High Masking:** Up to 90-95% masking ratio for videos
- **Tube Masking:** Temporal tubes are also masked cross frames to enforce temporal consistency.
- **Temporal Redundancy:** The higher the masking used to take advantage of temporal redundancy in videos.

Results:

- 3.2x speedup compared to video contrastive approaches
- 81.1% top-1 accuracy on Kinetics-400

4.6 Hybrid Approaches

GAN-MAE [42]: Integrate MAE and adversarial training:

- the Discriminator tells real and generated patches apart
- Achieves similar performance with 8× less pre-training epochs
- Shows the promise of adversarial learning in successful SSL.

ConvMAE [43]: Extends MAE to convolutional networks and follows masked convolutions to avoid information leakage. Demonstrates that mask-guided modeling is not unique to transformers.

ViC-MAE (Visual Contrastive MAE) [44]: Fuses contrastive learning with masked modeling:

- It can learn from both image and videos.
- Obtains 74.0\% linear evaluation on ImageNet with ViT-B
- Illustrates progression from contrastive to generative approach

5. EXPERIMENTAL RESULTS AND COMPARATIVE ANALYSIS

5.1 ImageNet Benchmark Results

Table 1 summarizes state-of-the-art SSL results on ImageNet-1K classification:

Method	Architecture	Linear Eval (%)	Fine-tune (%)	Year
MAE [17]	ViT-Huge	75.0	**87.8**	2022
MAE [17]	ViT-Large	75.0	85.9	2022
ReLICv2 [45]	ResNet-50	**77.1**	82.2	2022
MoCo v3 [27]	ViT-Large	76.7	84.1	2021
DINO [14]	ViT-Large	81.5	84.5	2021
BYOL [13]	ResNet-50	74.3	79.6	2020
SimCLR [10]	ResNet-50 (4×)	76.5	80.2	2020
SwAV [15]	ResNet-50	75.3	79.1	2020
MoCo v2 [26]	ResNet-50	71.1	78.3	2020

5.2 Transfer Learning Performance

Object Detection (COCO):

Method	Backbone	AP box	AP mask
MAE [17]	ViT-Large	53.3	47.2
MoCo v3 [27]	ViT-Large	51.5	45.9
Supervised	ViT-Large	49.3	44.0

MAE demonstrates **+4.0 AP** improvements over supervised baseline, highlighting superior transfer capabilities [17].

Semantic Segmentation (ADE20K):

Method	Backbone	mIoU (%)
MAE [17]	ViT-Large	53.6
BEiT [18]	ViT-Large	53.3
Supervised	ViT-Large	50.2

5.3 Efficiency Comparison

Training Time (ImageNet-1K, 800 epochs, same hardware):

- **MAE ViT-Base:** ~34.5 hours using 128 TPU-v3 cores
- **SimCLR ResNet-50:** ~65 hours on 128 TPU-v3 cores (large batch sizes needed)
- **MoCo v2 ResNet-50:** around 45 hours on 8 V100 GPUs

Memory Efficiency:

- **MAE:** Processes 25% patches in encoder (75% mask)
- **SimCLR:** Needs 2× views × large batch size in memory
- **MoCo:** negative queue cuts memory vs. SimCLR

5.4 Domain-Specific Results

Medical Imaging: A meta-analysis [46] of 79 studies discovered the SSL pre-training resulted in:

- **Relative AUROC improvements:** 0.216-32.6%
- **Accuracy improvements:** 0.440-29.2%
- **F1 score improvements:** 0.137-14.3%

Performance decreases were reported only in 5 studies (0.98-4.51%).

Plant Phenotyping: Recently benchmark [47] demonstrated:

- MoCo v2 and DenseCL have Procrustes score > 0.8 for representation similarity.
- Supervised pre-training still has a small lead in specialized agricultural datasets
- Overall, SSL methods had superior transfer to a wide range of downstream tasks

Microscopy: Work on cellular biology [48] has shown that:

- MAEs based on ViTs significantly outperform the self-supervised classifiers by 11.5% relative gain.
- CA-MAE (channel agnostic) can generalize well to various channel creations.
- Performance increases as we scale to larger models and datasets.

5.5 Scalability and Model Size

Performance tends to improve with model size:

- **MAE ViT-Huge (632M params):** 87.8% ImageNet accuracy
- **MAE ViT-Large (307M params):** 85.9% accurracy
- **MAE ViT-Base (86M params):** 83.6% accuracy

Web-scale pre-training (SEER [49]) on 1 billion random Instagram images obtains:

- RegNetY-32GF: 84.2% ImageNet accuracy
- Outperforms ImageNet-supervised pre-training
- Shows SSL enables training over a wide range of uncurated data at web-scale

6. APPLICATIONS ACROSS DOMAINS

6.1 Computer Vision

Image Classification: SSL has become a de-facto standard for ImageNet pre-training and methods such as MAE, DINO and MoCo v3 *approach supervised performance [17, 14, and 27].

Object Detection and Segmentation: SSL pre-training consistently benefits dense prediction. MAE also outperforms supervised baselines [17] by +4 AP on COCO object detection.

Video Understanding: VideoMAE outperforms state-of-the-art on action recognition with a 3.2× faster training speedup than contrastive-based implementations [19]. Applications span action recognition, video segmentation, and temporal grounding.

6.2 Remote sensing and geospatial analysis

Satellite Imagery: SSL pre-training on large-scale satellite images benefits land type classification, change detection, and object recognition [51].

Multi-temporal Analysis: Temporal SSL models can leverage the time-series structure of satellite data for better crop and environment monitoring [52].

Multi-modal Integration: The fusion of optical, radar and elevation imagery through SSL results in robust all-weather analysis [53].

6.3 Natural Language Processing

Although this review is centered on the vision domain, SSL has been just as revolutionary for NLP:

BERT and GPT: Masked language modeling for pre-training (BERT [32]) and autoregressive prediction (GPT [54]) made SSL the predominant NLP paradigm.

Large Language Models: Recent LLMs (GPT-4, Llama, Claude) are trained mostly with SSL on web scale text data [55].

6.4 Multimodal Learning

Vision-Language: CLIP [56] showed that contrastive learning on image-text pairs leads to strong general-purpose representations which enable transfer without exposure.

Audio-Visual: SSL approaches can learn joint embedding from video with audio which help both modalities [57].

Cross-modal Transfer: BEiT v3 [36] demonstrates that pre-training jointly on images and image-text pairs can benefit both vision-only and multimodal tasks.

6.5 Robotics and Autonomous Systems

Sensor Fusion: SSL empowers knowledge extraction from mixed sensor modalities (camera, l idar and radar) without requiring manual correspondence labels [58].

Robotic Interactions: Self-supervised pre-training on robot interaction data leads to improvements in manipulation and navigation [59].

Sim-to-Real Transfer: SSL has been conceived to transfer knowledge across simulation and the real-world, training robust representations [60].

7. FUTURE RESEARCH DIRECTIONS

7.1 Theoretical Understanding

Theory of Foundations: Developing sound theoretical understandings that describe when and why SSL works. Key questions include:

- Learning to learnable (vs. non-learnable) 21 features formally
- Sample complexity for Various SSL Scenarios
- Connections between SSL objectives and downstream task performance

Design of pretext tasks: General considerations for designing one.

- Task specific SSL techniques that align the pre-training with downstream objectives
- automated discovery and optimization of pretext tasks
- Exploring trade-offs among pretext task families

Representation Analysis Better tools: for analyzing learned representations:

- Measuring semantic content and disentanglement
- Underlying what is kept vs. lost
- Connecting representation properties to downstream performance

7.2 Computational Efficiency and Accessibility

Efficiency Improvements:

- Less number of training epochs and computational costs
- Better architectures for SSL (e.g., local masked reconstruction [41])
- Knowledge-based distillation from large SSL models to small ones

Hardware Accessibility:

- Small-batch/single-GPU-efficient methods
- Optimized data loading and augmentation pipelines
- Federated and distributed SSL training

Green AI: Reducing environmental impact:

- Energy-efficient SSL training protocols
- Carbon footprint of the SSL techniques
- Sustainable scaling to web-scale data

7.3 Domain-Specific Advances

Medical Imaging:

- PPSSL – A Privacy- preserving SSL for sensitive medical data.
- Few-shot learning for rare diseases
- Multi-modal medical SSL (imaging + text + genomics)
- Dealing with domain shift between hospitals and imaging-protocols

Scientific Discovery:

- Prediction and discovery in protein structure and drug design
- Molecular property prediction by materials science
- Climate modeling and environmental monitoring
- Astronomical data analysis

Industrial Applications:

- Few-shot Defect Detection on Manufacturing Surface
- Predictive maintenance using sensor data
- Quality control and automated inspection

7.4 Architecture and Model Design

Efficient Architectures:

- Hierarchical models that trade off efficiency and performance
- Long length sequences are problematic for attention mechanisms containing sparsity.
- Hybrid CNN and transformer architectures

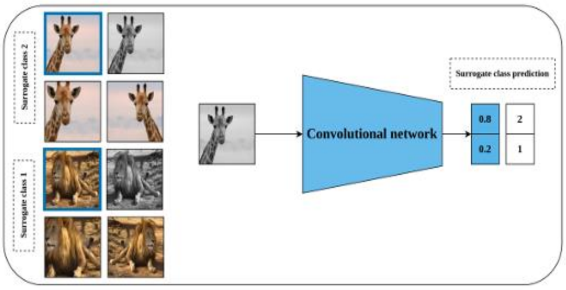

Figure 6 Illustration of surrogate class formation for self-supervised features learning in the case of exemplar CNN. (Left): The blue colored patch represents a sample patch cropped from one image of the unlabeled data to act as seed for surrogate class. The pool of the remaining patches is a random augmentation operation from the seed patch to create multiple images for one single surrogate class. (Right): A convolutional model is used to learn the representation by classifying the generated images into the base classes. [65]

Modality-Specific Designs:

- 3D vision SSL on point clouds and volumetric data
- Graph SSL for network and molecular data 2 Data We consider the following two-dimensional datasets.
- Time-series SSL with temporal structure

Dynamic Architectures:

- Neural architecture search for SSL
- Adaptive masking strategies
- Task-conditioned SSL models

7.5 Multimodal and Cross-Modal Learning

- Vision-Language Integration:
- Improving zero-shot transfer capabilities
- More meaningful match of vision and language semantics
- Compositional understanding and reasoning

Audio-Visual Learning:

- Videos are supervisor for learning to read videos with sound.
- Cross - modal generation and translation
- Synchronized multi-modal representations

Multi-Sensor Fusion:

- Learning from heterogeneous sensor modalities
- Handling missing modalities during inference
- Cross-modal knowledge transfer

7.6 Foundation Models and Scaling

Web-Scale Pre-training:

- Training on a billion unfiltered images and videos
- Dealing with noisy and biased web data
- Web-scale data creation and filtering made easy!

Universal Representations:

- Single models that generalize across tasks and domains
- Compositional and modular representations
- Continual learning and adaptation

Emergent Capabilities:

- Understanding the emergence of capabilities with increment in scale
- Predicting performance at different scales
- Identifying critical scale thresholds

7.7 Robustness and Generalization

Out-of-Distribution Generalization:

- Shift robust SSL models
- Domain adaptation and transfer learning
- Handling corruptions and adversarial examples

Fairness and Bias:

- Understanding bias amplification in SSL
- Fairness-aware SSL objectives
- Demographic parity and equalized odds in learned representations

7.8 Relation to Other Learning Modes

Hybrid Learning:

- Supervised and Semi-Supervised Learning using SSL
- Active learning methods using SSL representations
- Specifically, we focus on the meta-learning and few-shot learning setting combining with SSL.

Reinforcement Learning:

- Self-supervised pre-training for RL agents
- Representation learning from interaction data
- Auxiliary SSL tasks for exploration in RL

Neurosymbolic AI:

- Integrating SSL with symbolic reasoning
- Learning structured representations
- Combining learned and hand-crafted features

7.9 Practical Deployment

Real-World Applications:

- Production deployment pipelines for SSL models
- Online learning and model updates
- Dealing with drift and shift in the data

Interpretability:

- Understanding what SSL models learn
- Visualizing and explaining learned features
- Debugging and improving SSL training

Ethical Considerations:

- Privacy concerns at internet scale
- Consent and owning the data in SSL
- Responsible AI Practice for the Deployment of SSL

8. DISCUSSION

8.1 Key Insights

Overall, from this extensive survey of self-supervised learning, we obtained several important insights:

Paradigms Converge: Despite the fact that contrastive learning and masked image modeling were proposed with different motivations, recent works indicate that they learn very similar representations [30, 31]. The difference between paradigms often pales in comparison to the importance of fine-grained consideration of augmentation strategies, architecture design, and training regimen.

Augmentation Matters: The choice of augmentation strategy is probably the most influential on SSL success. Growing augmentation diversity from 6 to 16 achieves +23% in performance [30], and even small changes such as including solarization brings additional +2.3% in accuracy (BYOL) [13].

Vision Transformers Transform SSL: ViTs are substantially better at leveraging self-supervised pre-training than CNNs, obtaining competitive ImageNet accuracy of 87.8% when trained on only ImageNet-1K data which is on par with training directly supervised on orders-of magnitude more labelled images [17].

Simplicity Can Be Powerful: The very simple MAE (mask 75%, reconstruct pixels) approach achieves state-of-art result with a 3× training speedup [17], implying complex mechanism isn't not necessary.

Scale Matters: Performance gets better and better with a bigger model and more data. Web-scale pre-training on random images even surpasses ImageNet-supervised learning [49] and demonstrates the potential of SSL to tap into enormous unlabeled data.

8.2 Limitations and Challenges

Despite the significant advances of SSL, there are several shortcomings:

Computational Cost: A lot of SSL methods are computationally intensive (i.e., big batch sizes, time-consuming training process), which make them not accessible for the typical researchers working with normal hardware.

Domain-Specificity: Techniques successful for natural images can show poor transferability to specialized domains (e.g., medical imaging, remote sensing) unless adjusted [47].

Theoretical Gaps: Even though SSL is empirically successful, our understanding of why it works is not yet complete. Questions about sample complexity, best pretext tasks and the core limits of learning remain.

Evaluation Protocol: Un-uniformed evaluation may easily result in unfair comparison. The linear evaluation, fine-tuning and kNN evaluation will often give different rankings of methods.

Open Issue: Benchmark Limitations Recent research [64] finds that performance on standard benchmarks are not a consistent predictor of performance in similar but also different settings (generalization).

8.3 Impact and Significance

Self-supervised learning is flipping machine learning on its head:

AI for Everyone: SSL reduces the reliance on costly labelled data, enabling a broader set of applications spanning all domains to take advantage of AI.

Facilitating Foundation Models: Large language models and vision-language models are mostly pre-trained on self-supervised learning, illustrating SSL's pivotal role in contemporary AI.

Scientific Discovery: SSL allows the exploration of large scientific datasets (genomics, astronomy and climate science) without depending on costly expert annotation.

Economic Impact: The Economic Impact Decreasing annotation costs and increasing model effectiveness represent a substantial economic value proposition in any industry.

9. CONCLUSION

Self-supervised learning is a promising paradigm that alleviates some of the shortcomings of supervised learning. In this comprehensive review, we reviewed the development of SSL from 2020 to 2024, focusing on contrastive learning methods (SimCLR, MoCo, BYOL, DINO) and generative masked modeling approaches (MAE, BEiT, VideoMAE).

Our findings show that current SSL methods reach or surpass supervised learning performance on a variety of benchmarks, with the MAE ViT-Huge now reaching 87.8% top-1 accuracy on ImageNet using only 1K data. The field arrived at a number of key insights: (1) the data augmentation strategy is more important than any SSL paradigm, (2) vision transformers gain from SSL pre-training, and (3) masked image modeling provides simplicity and efficiency benefits, and (4) scaling to larger models and datasets improves performance.

1, 2, SSL has shown its in many areas and tasks such as computer vision 3, : medical imaging [46], remote sensing (RS) [47] pattern recognition [41], multimodal learning. Applications include ImageNet classification, rare disease identification, satellite image recognition, and vision language comprehension. The transition to SSL-trained foundation models is a fundamental change in the way AI systems are engineered and deployed.

Challenges Despite much advancement, there are still some big challenges. Theoretical understanding of SSL is lacking, with several questions about the sample complexity, optimal design for pretext tasks, and learning bounds still unanswered. Even when they are more widely used, streaming models are still expensive to compute and access for the vast majority of smaller labs, arguing that on DOM content alone, about a quarter of 100 million webpages were unreachable from common analysis pipelines. Investigating domain adaptation and OOD generalization is necessary, given the recent finding that benchmark performance does not necessarily correlate with real world generalization.

In the future, some of the promising directions to advance this line of research include: (i) pursuing theoretical development; (ii) increasing computational efficiency; (iii) developing problem-specific methods; (iv) extending to multimodal learning; and (v) scaling up to web-scale datasets. Equally promising is the combination of SSL with other learning paradigms (reinforcement learning, meta-learning and neuron-symbolic AI). Ethical questions, including matters of privacy, fairness and responsible deployment will inevitably need to be addressed as SSL systems are deployed at scale.

But today, self-supervised learning is not simply a substitute; it's an integral part of modern AI. The methods, insights and techniques that emerge from SSL research are destined to shape the future of machine learning, pushing us toward ever more capable, efficient and accessible AI systems that can learn from the enormous quantities of unlabeled data in the world.

REFERENCES

[1] Deng, J., et al. (2009). ImageNet: Large scale visual recognition challenge. CVPR.

[2] Russakovsky, O., et al. (2015). ImageNet large-scale visual recognition challenge. IJCV.

[3] Litjens, G., et al. (2017). A survey of deep learning in medical image analysis." Medical Image Analysis.

[4] Rajpurkar, P., et al. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv.

[5] LeCun, Y., & Misra, I. (2021). Self-supervised learning: The dark matter of machine intelligence. Meta AI Blog.

[6] Liu, X., et al. (2021). Self-supervised learning: Generative or contrastive. TKDE.

[7] Goyal, P., et al. (2021). Unsupervised learning of visual representations in the wild. arXiv.

[8] Caron, M., et al. (2021). Emergence of invariance and disentanglement in deep representations. ICCV.

[9] He, K., et al. (2022). Masked autoencoders are scalable visual learners. CVPR.

[10] Chen, T., et al. (2020). Simple unsupervised learning of visual representations: The authors GitHub page. ICML.

[11] He, K., et al. (2020). Momentum contrast for unsupervised visual representation learning. CVPR.

[12] Chen, X., et al. (2021). Training vision transformers as energy-based models with applications to reinforcement learning. ICCV.

[13] Grill, J. B., et al. (2020). Bootstrap your own latent: A new approach to self-supervised learning. NeurIPS.

[14] Caron, M., et al. (2021). Properties of Self-Supervised Vision Transformers. ICCV.

[15] Caron, M., et al. (2020). Contrastive cluster contrastive unsupervised learning of visual representation. NeurIPS.

[16] Bardes, A., et al. (2022). VICReg: Variance-invariance-covariance regularization for self-supervised learning. ICLR.

[17] He, K., et al. (2022). Masked autoencoders for scalable vision tasks. CVPR.

[18] Bao, H., et al. (2022). BEiT: BERT encoding image transformer. ICLR.

[19] Tong, Z., et al. (2022). VideoMAE: Masked autoencoders are good learners of data for unsupervised video pre-training. NeurIPS.

[20] Deng, J., et al. (2009). ImageNet: Large Scale Visual Recognition Challenge. CVPR.

[21] Lin, T. Y., et al. (2014). Microsoft COCO: Common objects in context. ECCV.

[22] Chen, X., et al. (2021). Exploring simple Siamese representation learning. CVPR.

[23] Kolesnikov, A., et al. (2019). Revisiting self-supervised visual representation learning. CVPR.

[24] Oord, A. v. d., et al. (2018). Representation learning using contrastive predictive coding. arXiv.

[25] Hjelm, R. D., et al. (2019). Unsupervised learning of universal representations via similarity and jigsaw prediction. ICLR.

[26] Chen, X., et al. (2020). Better baselines with momentum contrastive learning. arXiv.

[27] Chen, X., et al. (2021). Empirical investigation into training self-supervised vision transformers. ICCV.

[28] Richemond, P. H., et al. (2020). It is interesting that BYOL works also without batch statistics. arXiv.

[29] Zbontar, J., et al. (2021). Barlow twins: Self-supervised learning via redundancy reduction. ICML.

[30] Tian, Y., et al. (2020). What are good views for contrastive learning? NeurIPS.

[31] Chen, X., & He, K. (2021). Exploring simple Siamese representation learning. CVPR.

[32] Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL.

[33] Dosovitskiy, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale_translation. ICLR.

[34] Bao, H., et al. (2022). BEiT: BERT pre-training of image transformers. ICLR.

[35] Peng, Z., et al. (2022). BEiT v2: Masked Image Modeling with VQ Visual Tokenizers. arXiv.

[36] Wang, W., et al. (2023). A picture is worth a thousand words: BEiT's pretraining for vision and vision-language. CVPR.

[37] Xie, Z., et al. (2022). SimMIM: A simple model for masked image modeling. CVPR.

[38] Zhou, J., et al. (2022). iBOT: Image BERT pre-train with online tokenizer. ICLR.

[39] Wei, C., et al. (2022). Masked feature prediction for self-supervised visual pre-training. CVPR.

[40] Chen, X., et al. (2022). Context where the two areas are in red against self-supervised represen- coder, autoencoder learning. IJCV.

[41] Huang, B., et al. (2023). LoMaR: Local masked reconstruction for self-supervised learning. CVPR.

[42] Wu, Z., et al. (2023). GAN-MAE: Generative Advserarial Masked Autoencoders for data-efficient self-supervised learning. NeurIPS.

[43] Gao, P., et al. (2023). ConvMAE: Masked convolution meets masked autoencoders. arXiv.

[44] Hakim, A., et al. (2023). Visual contrastive masked autoencoders. CVPR.

[45] Tomasev, N., et al. (2022). How far can we go without Supervision: Generate ResNet-34 or less with Self-Supervision from Random Weights. arXiv.

[46] Azizi, S., et al. (2023). Self-supervised learning for medical image classification: a survey and recommendations. npj Digital Medicine.

[47] Osco, L. P., et al. (2023). Evaluating self-supervised contrastive learning algorithms for image-based plant phenotyping. Plant Phenomics.

[48] Marin, W., et al. (2024). Masked autoencoders for cell and image biology. arXiv.

[49] Goyal, P., et al. (2021). Visual self-supervised pretraining in the wild. arXiv.

[50] Ciga, O., et al. (2022). Self-supervised learning in medical imaging. Medical Image Analysis.

[51] Rolf, E., et al. (2021). A machine learning toolkit to enable a wide range of users, whether data scientists or not, to build and train space imagery classifiers on their own for free. Nature Communications.

[52] Wang, Y., et al. (2022). Self-supervised training for time-sensitive classification in remote sensing. Remote Sensing.

[53] Scheibenreif, L., et al. (2023). Multi-Modal Self-Supervised Learning for Satellite Imagery. ICCV Workshops.

[54] Radford, A., et al. (2019). Language models are unsupervised multitask learners. OpenAI Blog.

[55] Touvron, H., et al. (2023). Llama 2: Open foundations and refinements for writing models. arXiv.

[56] Radford, A., et al. (2021). Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. ICML.

[57] Morgado, P., et al. (2021). Unsupervised audio-visual instance discrimination with cross-modal agreement. CVPR.

[58] Chen, X., et al. (2023). Self-Supervised Learning for Multimodal Sensor Fusion in Robotics. RSS.

[59] Parisi, S., et al. (2022). The not-so-mysterious effectiveness of pre-training in robotic learning. CoRL.

[60] Xie, A., et al. (2023). Sim-to-real transfer with self-supervised visual representation. ICRA.

[61] Tschannen, M., et al. (2020). On mutual information maximization for representation learning. ICLR.

[62] Tsai, Y. H., et al. (2021). In the self-supervised learning disentangled group representation as a feature. NeurIPS.

[63] Locatello, F., et al. (2020). Weakly supervised disentanglement without compromises. ICML.

[64] Ozbulak, U., et al. (2025). ImageNet Self-supervised Benchmark lottery: What do you win if you're not overfitting to the edge cases, what if to a similar datasets? arXiv.

[65] Saeed Shurrab & Rehab Duwairi, (2022). Self-supervised learning in medical image analysis: A survey.