Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

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

Hybrid Learning Paradigms: Bridging Supervised and Unsupervised Methods for Semi-Supervised Deep Learning

Virendra Tank
Shri Mahaveer College, Jaipur, India
0009-0003-9126-0982

Abstract - The ever-increasing size of unlabelled data and the expensive cost for manual annotation have driven research on semi-supervised learning (SSL), which aims to take full advantage of labelled and unlabeled samples. There are various hybrid learning approaches that connect supervised and unsupervised counterparts to each other in the contemporary era, including self-supervised pretraining paradigms, consistency regularization methods, and pseudo-labeling practices. In this paper, we study three popular backbone architectures SimCLR, BYOL and MoCo to investigate their specific character is- tics, understand their theoretical roots and applications in both computer vision and natural language processing. The coming together of these two methods is a breakthrough in deep learning research, allowing researchers and developers to achieve almost-supervised performance with significantly less labels.

Keywords
Semi-supervised learning, Self-supervised learning, Contrastive learning, Consistency regularization, Pseudo-labeling, Deep learning

1. INTRODUCTION

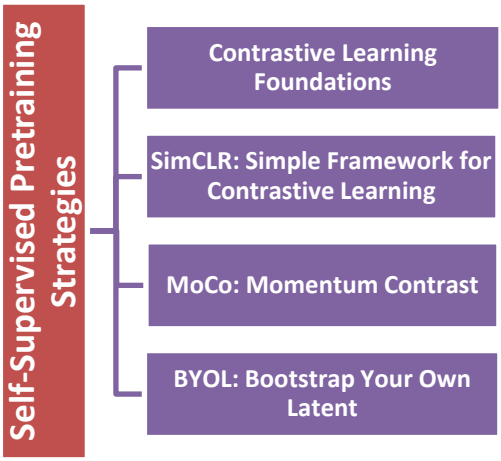
Deep learning has transformed the field of artificial intelligence by learning multi-level representations directly from data [1]. But, the performance of supervised learning approaches fundamentally rely on plentiful labeled data, which is expensive (e.g., salaries for experts), time-consuming (intuitive labeling), and sometimes barely accessible. The ImageNet dataset, for example, took more than 25,000 workers and several years to annotate [2]. On the other hand, there is plenty of unlabeled data in almost all fields. This basic asymmetry has stimulated the development of semi-supervised learning methods that attempt to exploit the best properties of supervised and unsupervised learning [3, 4].

Semi-supervised learning resides between the extreme cases of supervised learning, where there is a label for each training example, and unsupervised learning, where no labels are employed. The basic idea is that unlabeled data can convey useful structural information on the underlying data distribution, and thereby boost model performance beyond what can be achieved with a limited amount of labeled data [5]. Recent techniques within self-supervised learning, contrastive methods and consistency-based approaches have radically increased the effectiveness of semi-supervised methods, especially when labeled data is limited [6].

In this review, we seek to integrate recent advances in hybrid learning methods that can leverage both supervised and unsupervised signals. We focus on the following three main methodological families: self-supervised pretraining approaches to learn representations without labels, consistency efficiency-focused regularization methods for inducing invariance properties, and pseudo-labeling techniques that give birth iteratively a labeled dataset. Throughout, we underscore both theoretical underpinnings and practical application on various domains such as computer vision and natural language processing.

2. SELF-SUPERVISED PRETRAINING STRATEGIES

Self-supervised learning has been shown to be an effective paradigm for learning informative representations from unlabeled data by designing pretext tasks that are constructed from auxiliary information available in the data directly [6, 7]. Contrary to standard unsupervised models that concentrate on density estimation or clustering, self-supervised models learn features before fine-tuning them on labelled tasks. This two-stage procedure has proved to be highly effective, and frequently outperforms supervised learning or semi-supervised techniques using much smaller amounts of labeled data [8].



2.1 Contrastive Learning Foundations

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

Many state-of-the-art self-supervised approaches are grounded on contrastive learning [9], initially proposed in the context of metric learning. The idea is to discover a representation that encourages consistency of differently augmented views of the same example and inconsistency across examples [10]. This method directly optimizes the model to possess invariances of data augmentations with discriminative representations which captures semantic similarity without any explicit labelling.

The contrastive learning objective can be cast into the InfoNCE loss that maximizes the mutual information of representations for positive pairs and minimizes it for negative pairs [10]. The model is given a query representation and one positive key as well as K negative keys, and it learns to differentiate the positive key from the set of K+1 samples. This provides a formulation that links contrastive learning to noise-contrastive estimation and a theoretical grounding of why these methods work in learning meaningful representations [11].

2.2 SimCLR: Simple Framework for Contrastive Learning

SimCLR presented a simplified contrastive learning strategy, and obtained the state-of-the-art performance by designing sophisticated strategies of augmentation and architectural components [8]. The method works by generating multiple augmented views of the same image, which it treats as positive pairs, and using views from different images in the same batch as negative examples. SimCLR applies a combination of data augmentation operations such as random cropping, color distortion, and Gaussian blur to create different views.

A key architectural innovation in SimCLR is the projection head: a non-linear transformation before computing the contrastive loss. Interestingly, this projection head increases quality of representation despite being thrown away after pretraining, indicating that the contrastive objective benefits from a distinct feature space compared to downstream tasks [8]. Furthermore, SimCLR showed that increasing batch sizes not only increases the number of negatives per positive pair but it also improves performance a lot. Given the appropriate training configurations, SimCLR attained 76.5% top-1 accuracy on ImageNet with linear evaluation, challenging other supervised pretraining works.

2.3 MoCo: Momentum Contrast

MoCo copes with the computational difficulties of storing large negative sample sets in contrastive learning by using a neat-queue method [12]. Instead of using only large batch sizes, MoCo stores a dictionary of encoded representations as a dynamic queue. This queue acts as a big, consistent set of negative examples that is effective across large training iterations, and separates the number of negative samples from the batch size.

To maintain the stability of the representations in queue, MoCo uses momentum encoders (momentum encode updates the value at time t as an exponential moving average with a decay constant that is dependent on m) to update the query encoder $\theta\_key = m \cdot \theta\_key + (1-m) \cdot \theta\_query$, where $m \in [0, 1]$ is a coefficient which controls momentum [12]. This momentum based update allows avoiding sudden changes in the encoder that would make old queue entries incompatible with the current representations. MoCo v2, which introduced many improvements over SimCLR such as the MLP pro-jection head and stronger augmentations, scored 71.1% on ImageNet top-1 accuracy with linear evaluation, showing effectiveness of the momentum based approach [13].

2.4 BYOL: Bootstrap Your Own Latent

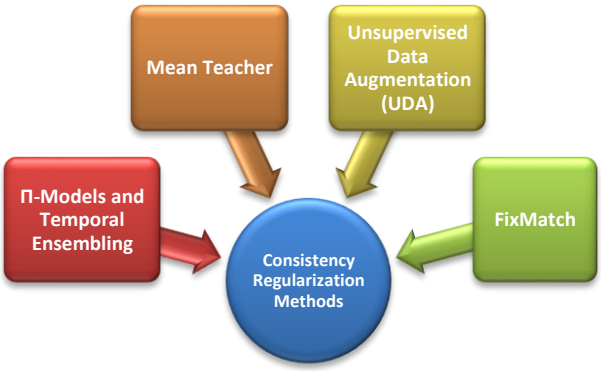
BYOL is a radical change from classic contrastive work as it abandons negative pairs entirely [14]. This method questioned the common sense that contrastive learning must come with negative samples to avoid collapsing of representation. BYOL relies on two neural networks: an online one that is actively trained via gradient descent, and a target network whose parameters are updated as exponential moving averages of the online counterpart.

At training time, the online network predicts the representation that would be produced by the target network for various augmented views of a given image. The asymmetric design with a prediction head of the target Siamese online network avoids collapsing to trivial solutions, [14]. BYOL attained 74.3% top-1 accuracy on ImageNet with linear evaluation, showing that self-supervision can be obtained through from pure prediction objectives without contrastive terms. Recent theoretical findings that the unsupervised loss of BYOL operates as implicit contrastive learning given batch normalization and predictor asymmetry [15].

3. CONSISTENCY REGULARIZATION METHODS

Consistency regularization is one of the prototypes semi-supervised learning methods that is built upon the smoothness assumption: “the predictions of a model should not change much under small perturbations to example or model weights” [16]. This is consistent with the manifold hypothesis, that high dimensional data lies on lower dimensional manifolds and points in the same manifold should have similar predictions [3].

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



3.1 Π-Models and Temporal Ensembling

In [17], the Π-Model introduced input-based consistency regularization as it generates predictions on two randomly augmented versions of each input and calibrates its parameters so that they are closer in terms of their discrepancy. The consistency loss, which is often based on mean squared error or KL divergence, is added to the loss on labeled examples. This simple trick greatly enhances learning on small labeled datasets by harnessing the power of unlabeled data and consistency enforcement over the output predictions.

Temporal Ensembling improves on this by updating EMA of predictions throughout the training epochs [17]. Instead of predicting twice per example in every epoch, temporal ensembling uses past predictions as a consistency target: $Z_i = \alpha Z_i + (1-\alpha)\, z_i$ where Z_i is the EMA prediction for example i, and z_i is the current prediction. This mitigates computational cost and stabilizes training by applying temporal smoothing on predictions, and results in better performance on CIFAR-10 and SVHN benchmarks.

3.2 Mean Teacher

The method extends the consistency regularization by using a teacher-student framework, in which the teacher model is a temporally averaged version of the student model [18]. The student is optimized to minimize both the supervised loss on labeled data but also a consistency loss between its predictions and the teacher's predictions on unlabeled data. Importantly, the teacher's parameters are updated using an exponential moving average of those of the student instead of gradient descent: $\theta_{teacher} = \alpha\theta_{teacher} + (1-\alpha)\theta_{student}$.

We mitigate the noise and instability of premature gradients that existing methods suffered from by offering more reliable consistency targets due to temporal averaging [18]. Mean Teacher outperforms temporal ensembling by a large margin, achieving the error rate of 6.28% on CIFAR-10 with only 4K labels, as compared to 12.16% when using temporal ensembling alone. The success of the method is based on the fact that teacher model delivers a smooth and accurate prediction, rather than predictions from single epoch, which actually computes self-ensemble without any computational overhead at inference.

3.3 Unsupervised Data Augmentation (UDA)

UDA blends the consistency regularization and complex augmentation strategies that are appropriate for domains [19]. In computer vision, UDA uses RandAugment which automatically chooses augmentations and then applies sequences of transforms with different strengths [20]. For NLP, UDA proposes new augmentation methods such as back translation (translating a sentence to another language and then back to the original), word replacement by TF-IDF-based similarity with contextualized embedding.

The consistency loss in UDA is calculated only on the unlabeled samples where the model makes high-confidence predictions on their original (not augmented) input, which inherently integrates confidence-based filtering with consistency enforcement [19]. This selective behavior avoids the model from learning improper invariances on uncertain predictions. UDA advanced the state of the art on several NLP benchmarks, achieving a 4.20% rate of errors in sentiment classification for IMDb with merely 20 labeled examples per class and escape human performance at.50% using domain specific augmentation strategies to guide consistent regularization.

3.4 FixMatch

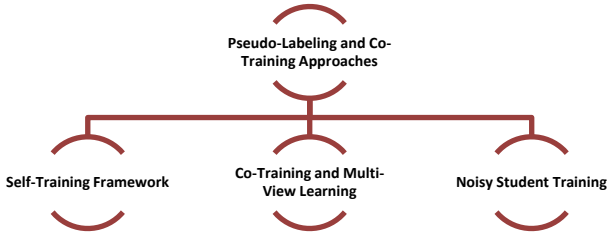
FixMatch combines consistency regularization with pseudo-labeling via an elegant weak to strong consistency principle [21]. The approach uses a weak augmentation (common flips and crops) for pseudo-labels and then employs strong augmentation (such as Rand Augment or CT Augment) to get the consistency target. Pseudo-labels are produced based on predictions with high confidence only for weakly augmented input and serve as targets for training with strongly augmented versions of the same input.

The joint objective is: $L = L_s + \lambda\, u \cdot 1(\max(q\ b) \geq \tau) \cdot H(\hat{q}\_b, p\_b)$, where L_s is the supervised loss, q b and p b is model's predictions the weakly and strongly augmented input [21]. This formulation incentivizes the model to make confident and uniform (consistent across runs or crops) predictions in spite of strong augmentations. On CIFAR-10, FixMatch reaches an accuracy of 94.93% with only 250 labels, showing that our approach is highly data efficient thanks to the synergy of consistency regularization and pseudo-labeling.

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

4. PSEUDO-LABELING AND CO-TRAINING APPROACHES

Pseudo-labeling is conceptually different to semi-supervised learning, and directly tackles label scarcity by synthesizing labels for unlabeled data [22]. Predictions by the model become pseudo labels for unlabeled instances, which are used to grow the training data and develop the model further by means of a variant of self-training.



4.1 Self-Training Framework

The self-training framework, which has its roots in early work by Scudder [23] and Yarowsky [24], iteratively produces pseudo-labels for unlabeled data based on the model's predictions at test time followed by training on all original labeled instances and newly generated pseudo-labeled data. Recent works [22] include a confidence thresholding, keeping pseudo-labels only when the model is highly-confident: ŷ _i = argmax_c p model(y=c|x_i) if max c p model (y= c |x i) ≥ τ.

The success of self-training depends on the quality of pseudo-labels and how to avoid confirmation bias, in which poor pseudo-labels can reinforce errors made by a model [25]. Methods such as label sharpening that transform weak probability distributions into hard one-hot labels are widely used for disparate loss smoothing methods to provide more clear learning targets [26]. Recent work analytically shows that ST (self-training) is effective when the learned decision boundary of a model is well-calibrated for high-confident regions such as those derived from informally labeled data.

4.2 Co-Training and Multi-View Learning

Co-training generalizes self-training by training multiple models on distinct views of the data, and enabling each model to produce pseudo-labels for the other model [27]. In the traditional co-training approach, it is assumed that features can be divided into two conditionally independent views on the basis of class label. Given this assumption, contrary predictions of models trained on alternate views can offer informative signals about prediction uncertainty.

Contemporary approaches relax the assumption of conditional independence and achieve diversity with separate initializations, architectures, or training strategies but not by an explicit feature partitioning [28]. Deep co-training trains multiple networks using varying architectures or initializations and make use of model diversity for more robust pseudo-labels than single-model methods. Apparently, when models with different inductive biases can reach consensus on their predictions for unlabeled examples, it is much more likely that those predictions are correct and the errors do not compound during iterative training.

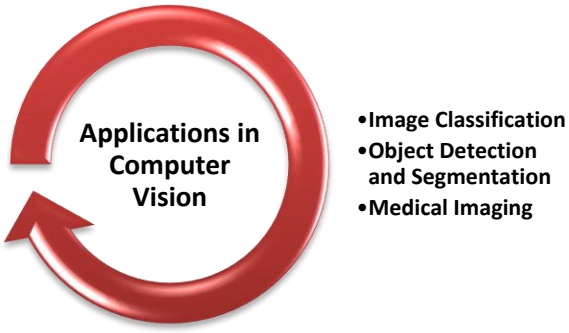
4.3 Noisy Student Training

Noisy Student is a powerful instance of a self-training technique at scale that surpasses state-of-the-art performance on ImageNet classification [29]. These approaches consists in training a teacher on labeled data and then use it to predict pseudo-labels for unlabeled samples before finally undertaking the training of a wider student over the union set with additional noise from data augmentation, dropout and stochastic depth.

The student model is intentionally larger than the teacher so that we have some extra capacity for learning from the pseudo-labeled data and capturing some patterns that the teacher may miss [29]. Noise is injected during the training of the student model which avoids the plaguing issue of students over fitting on teacher's predictions and encourages learning tougher features. This cycle can repeat with the student being turned teacher for the next round. Our Noisy Student model obtained 88.4% top-1 accuracy on ImageNet, setting a new state-of-the-art result and showing that iterative self-training can effectively utilize sets of unlabeled data with careful architectural and algorithmic changes.

5. APPLICATIONS IN COMPUTER VISION

Semi-supervised learning methods have been applied widely in the field of computer vision, where pixel-level or instance-level labeling is especially costly and time-consuming [30].



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

5.1 Image Classification

Semi-supervised approaches have excelled in several image classification benchmarks. On CIFAR-10, FixMatch matched the performance of fully supervised training, achieving 95.7% accuracy with just 250 labeled examples (40 per class) comparing to 96.1% [21] on the same task and with access to all 50k labels in the dataset. Self-supervised pretraining and fine-tuning has become a common practice and methods, such as SimCLR or MoCo, give strong initialization for downstream tasks [31]. The model and method of contrastive pretraining coupled with semi-supervised fine-tuning is a powerful paradigm that provides computational advantages to achieve competitive performances, given small labelled data.

5.2 Object Detection and Segmentation

In semi-supervised learning tasks these are even more difficult in dense prediction problems, because of structured outputs. Recent 'Consistency Regularization' and pseudo-labeling methods have been adapted to object detection by merely enforcing the consistency of bounding box predictions across augmented images [32, 33]. For semantic segmentation, MixMatch and FixMatch have been generalized to produce pixel-wise pseudo-labels without additional annotations with competitive results [30].

Self-supervised pretraining has achieved great success, particularly in segmentation tasks. Contrastively pretrained models on large unlabeled datasets learn representations that encode object boundaries and semantic structure, and transfer effectively to pixel-level prediction with little labeled data [34]. In medical imaging segmentation, using few expert annotations, self-supervised pretraining with consistency regularization has made it possible to apply deep learning models in clinical practice [35].

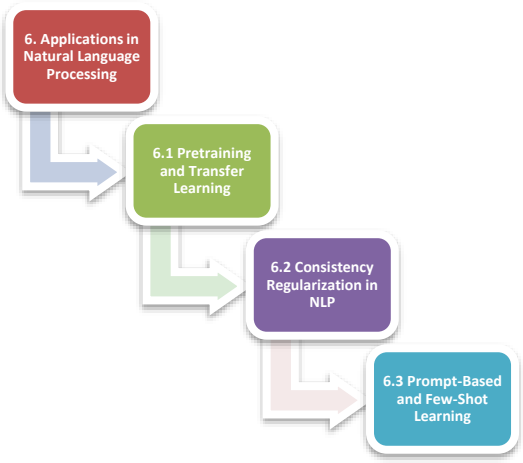
5.3 Medical Imaging

Medical imaging is a potentiality-rich application domain for which semi-supervised learning can alleviate the bottleneck of annotations. Annotation of radiological images is knowledge-dependent and time-consuming, making annotated data highly limited [36]. Pretraining have been widely used as standard procedure for pretraining on very large non-medical labeled data followed by fine-tuning on smaller medical imaging dataset such as tumor detection, organ segmentation and disease classification [37, 35].

Special regularization techniques for medical imaging adapted to the continuous-discrete nature of the prediction space are enforced predictions to remain robust under clinically-important transformations, they enhance robustness and generalizability [38]. The introduction of domain-specific augmentations, contrastive pretraining and consistency regularization has allowed semi-supervised approaches to approximate supervised performance with 10-20% fewer labels, democratizing medical AI applications.

6. APPLICATIONS IN NATURAL LANGUAGE PROCESSING

A number of groundbreaking developments have occurred with natural language processing through the use of semi-supervised and self-supervised learning, radically altering the way that models are built for language [39, 40].



6.1 Pretraining and Transfer Learning

The prevalence in modern NLP research is pretraining models on large text corpora and then fine-tuning them on a downstream task with small amounts of labeled data. BERT introduced masked language modeling, in which random tokens are masked and the model is trained to predict those tokens from their contexts [39]. Autoregressive Language Modeling: The GPT architectures are based on autoregressive language modeling, in which the next token is predicted conditionally on the preceding tokens [41, 40].

These pretraining objectives empower models to capture rich contextual representations of the text, encoding syntax and semantics, as well as world knowledge. Transfer learning from pretraining a model on downstream tasks with virtually no labeled data has become common and BERT, in particular, has been able to achieve state-of-the-art results across many NLP benchmarks by fine-tuning on task-specific examples [39]. The paradigm showcases the capacity of self-supervised learning at scale, for which models trained on billions of unlabeled tokens develop broad capabilities that are transferable to many tasks.

6.2 Consistency Regularization in NLP

UDA validated that the consistency regularization is effective for text and sequence tasks and they proposed domain-specific augmentation methods [19]. Back-translation can produce paraphrases by translating through an intermediate language and back, providing topic-

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

equivalent variations on labels. The TF-IDF-based word replacement by contextualized embedding from pre-trained LMs is another augmentation strategy that preserves the meaning.

The collaboration between pre-trained language models and consistency regularization pushes the frontier of low-resource NLP. On IMDb sentiment classification, UDA obtained 95.8% accuracy using only 20 labeled examples versus 88.7% for the standard supervised learning [19]. These illustrate the strong synergism between consistency-based semi-supervised learning and pretrained representations, which itself enables few-shot learning.

6.3 Prompt-Based and Few-Shot Learning

Massive language model pretraining has made few-shot and zero-shot learning feasible via prompt-based approaches [40, 42]. Pretrained models can be conditioned on well-crafted prompts to complete a wide variety of tasks with little or no task-specific fine-tuning from practitioners. GPT-3, conditioned with task descriptions or examples as input to its prefix [40], has shown impressive few-shot learning on a variety of tasks based on 300B tokens of pre-training.

This is the logical conclusion of semi-supervised learning in which massive self-supervised pretraining supplies overall competence, and a small number of labeled examples or prompts supervise task-specific behavior. Recent work in prompt design and in-context learning is investigating how to better elicit knowledge from language models, making labeled data even less necessary [43].

7. CHALLENGES AND FUTURE DIRECTIONS

Despite this, there are still several challenges. Many of these methods are sensitive to hyper-parameters, augmentation policies and discrepancy between labeled and the unlabeled data distribution [44]. In order to understand such insight, the when and why semi-supervised work is successful would require theoretical insight into the inductive bias of which they introduce [45].

Potential future directions would be devising more robust methods to cope with distribution shift [46], better confidence calibration given pseudo-labeling [47] and learning adaptive augmentation strategies [20]. The great computational burden of large scale self-supervised pretraining has led to efforts towards efficient training techniques [48]. Furthermore, combining several semi-supervised approaches via meta-learning would allow for automatic selection of the method in function of the dataset [49].

8. CONCLUSION

Hybrid learning frameworks that across supervised and unsupervised methods have revolutionized deep learning, etc. has led to state-of-the-art performance on several tasks despite using a small amount of labeled data. Self-supervised pretraining methods such as SimCLR [8], BYOL [14] & MoCo [12] learn strong representations from unlabeled data. Consistency regularization methods enforce invariances to improve the generalization [18, 21]. Pseudo-labeling methods repeatedly enlarge their training sets with the help of confident predictions [22; 29].

The confluence of these methods is indicative of a trend in machine learning. Instead of relying on labeled data as the only supervisory signal, recent methods highlight data structure, augmentation invariances and model consistency. This holistic view allows for the power of vast amounts of weakly labeled examples when the labels themselves are scarce. As the methodology matures and its theory becomes richer, semi supervision will be at the core of many practical instantiations and can finally make place for tasks where labels are scarce by nature or too expensive to obtain.

REFERENCES

[1] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
[2] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K. & Fei-Fei, L. (2009). ImageNet: A larg e-scale hierarc hical image database . CVPR, 248-255.
[3] Chapelle, O., Schölkopf, B., & Zien, A. (2006). Semi-supervised learning. MIT Press.
[4] Van Engelen, J. E., & Hoos,H. H. (2020). A survey on semi-supervised learning. Machine Learning, 109(2), 373-440.
[5] Zhu, X. (2005). Semi-supervised learning literature survey. 1530, University of Wisconsin-Madison.
[6] Jing, L., & Tian, Y. (2021). Self-Supervised Visual Feature Learning with Deep Neural Networks: A Survey. TPAMI, 43(11), 4037-4058.
[7] Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J et al. (2021) Self-supervised learning: Generative or contrastive. TKDE, 35(1), 857-876.
[8] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). Simclr: A simple framework for contrastive learning of visual representations. ICML, 1597-1607.
[9] Hadsell, R., Chopra, S., & LeCun, Y. (2006). Unsupervised feature learning by learning an invariant mapping. CVPR, 2, 1735-1742.
[10] Oord, A., Li, Y., and Vinyals, O. (2018). Unsupervised representation learning with contrastive predictive coding. arXiv:1807.03748.
[11] Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., & Saunshi, N. (2019). A theoretical analysis of contrastive unsupervised representation learning Abhijeet SrivastavaUniversity of Michigang. ICML, 5628-5637.
[12] He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. CVPR, 9729-9738.
[13] Chen, X., Fan, H., Girshick, R., & He, K. (2020). Better baselines with momentum contrastive learning. arXiv:2003.04297.
[14] Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E et al. (2020). Bootstrapping your own latent: A new approach to self-supervised learning. NeurIPS, 33, 21271-21284.
[15] Tian, Y., Chen, X., Ganguli, S. (2021). Understanding self-supervised learning dynamics without contrastive pairs. ICML, 10268-10278.

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

[16] Sajjadi, M., Javanmardi, M., & Tasdizen, T. (2016). On training robust deep semi-supervised learning models with stochastic transformations and re-labeling. NeurIPS, 29, 1163-1171.

[17] Laine, S., & Aila, T. (2017). Temporal ensembling for semi-supervised learning. ICLR.

[18] Tarvainen, A., & Valpola, H. (2017). Better teacher better dropout: Weight-averaged consistency targets improve semi-supervised deep learning. NeurIPS, 30, 1195-1204.

[19] Xie, Q., Dai, Z., Hovy, E., Luong, T., and Le, Q. 2020. Unsupervised data augmentation for consistency training. NeurIPS, 33, 6256-6268.

[20] Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. (2019). RandAugment: Practical automated data augmentation. CVPRW, 3008-3017.

[21] Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk E. D., Kurakin A. & Li C. L (2020). FixMatch: Simplifying semi-supervised learning with consistency and confidence. NeurIPS, 33, 596-608.

[22] Lee, D. H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural network. ICML Workshop, 3(2), 896.

[23] Scudder, H. (1965). Some adaptive pattern recognition machines [Research Report]. IEEE Trans. Information Theory, 11(3), 363-371.

[24] Yarowsky, D. (1995). Unsupervised word sense disambiguation with arbitrary number of senses. ACL, 189-196.

[25] Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., & McGuinness K. (2020). Pseudolabeling and confirmation bias in deep semi-supervised learning. IJCNN, 1-8.

[26] Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A. Raff el, C. (2019). MixMatch: A holistic approach to semi-{supervised} learning. NeurIPS, 32, 5049-5059.

[27] Blum, A., & Mitchell, T. (1998). Using labeled and unlabeled data in co-training. COLT, 92-100.

[28] Qiao, S., Shen, W., Zhang, Z., Wang, B., & Yuille, A. (2018). Deep co-training for semi-supervised image recognition. ECCV, 135-152.

[29] Xie, Q., Luong, M. T., Hovy, E., & Le, Q. V. (2020). Unsupervised learning Self-training with Noisy Student improves ImageNet classification. CVPR, 10687-10698.

[30] Ouali, Y., Hudelot, C., and Tami, M. (2020). A survey of deep semi-supervised learning. arXiv:2006.05278.

[31] Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. (2020). Large self-supervised models are powerful semi-supervised learners. NeurIPS, 33, 22243-22255.

[32] Jeong, J., Lee, S., Kim, J., Kwak, N. (2019). Object detection with a contrastive loss over code words. NeurIPS, 32, 10759-10768.

[33] Sohn, K., Zhang, Z., Li, C. L., Zhang, H., Lee, C. Y., & Pfister, T. (2020). Easy to learn semi-supervised object detection. arXiv:2005.04757.

[34] X. Wang, R. Zhang, C. Shen, T. Kong and L. Li,papertitle=Single-Guided Image Generation by TransformersXXXtitle=papersingle-guidedX[papersingle-guidedX], year=2021. Thick contrastive learning for visual pre-training self-supervisedly. CVPR, 3024-3033.

[35] Chaitanya, K., Erdil, E., Karani, N., & Konukoglu, E. (2020). Global and local contrastive learning for medical image segmentation with limited annotations. NeurIPS, 33, 12546-12558.

[36] Zhou, S. K., Greenspan, H., Davatzikos, C., Duncan, J. S., Van Ginneken, B., Madabhushi, A et al. (2019). Deep learning in medical imaging: overview and future 85 1 Introduction With the significant advances in big data and computing technolo- gies, deep learning has been developing rapidly. Proc. IEEE, 109(5), 820-838.

[37] Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G. et al. (2017). A semi-supervised approach for knowledge-driven network-based cardiac MR image segmentation. MICCAI, 253-260.

[38] *Bortsova, G., Dubost, F., Hogeweg, L., Katramados, I. and de Bruijne, M. (2019). Robust semi-supervised medical image segmentation through learning transforms under transformations. MICCAI, 810-818.

[39] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL, 4171-4186.

[40] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P. et al. (2020). Language models are few-shot learners. NeurIPS, 33, 1877-1901.

[41] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.

[42] Schick, T., & Schütze, H. (2021). Leveraging cloze-questions for few-shot text classification and natural language inference. EACL, 255-269.

[43] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi,H., & Neubig,G.(2023). Pre-train, prompt and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), 1-35.

[44] Oliver, A., Odena, A., Raffel, C., Cubuk, E. D., and Goodfellow, I. J. (2018). Reasonable assessment of deep semi-supervised learning techniques. NeurIPS, 31, 3235-3246.

[45] Balcan, M. F., & Blum, A.. (2010). From this to that using focal losses: Online calibration of black-box models. JACM, 57(3), 1-46.

[46] Yang, X., Song, Z., King, I. & Xu, Z. (2021). A survey of deep semi supervised learning. arXiv:2103.00550.

[47] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). Calibrating modern neural networks. ICML, 1321-1330.

[48] Chen, X., & He, K. (2021). Exploring simple Siamese representation learning. CVPR, 15750-15758.

[49] Hospedales, T., Antoniou, A., Micaelli, P., and Storkey, A. 2021. Meta-learning in artificial neural networks: A survey. TPAMI, 44(9), 5149-5169.