SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I7P108

Volume 12 Issue 7, 90-101, July 2025

Original Article

Enhancing Image Classification Performance through Hybrid Self-Supervised Learning Strategies

Deepa S[1], Sheetal[2], Alli A[3], Rashmi Siddalingappa[4]

[1,4]*Department of Computer Science, Christ University, Bengaluru, India.*
[2,3]*Department of Computer Applications, Presidency College, Bengaluru, India.*

[1]*Corresponding Author : sdeepa369@gmail.com*

Received: 04 May 2025 Revised: 06 June 2025 Accepted: 07 July 2025 Published: 31 July 2025

Abstract *- Image classification is a cornerstone of computer vision, with the applications spanning healthcare, autonomous driving and security. The dependence on large labeled datasets for supervised learning poses significant challenges, particularly in specialized fields where the labeled data is scarce and expensive to obtain. Self-supervised learning (SSL) has emerged as a promising paradigm, enabling models to learn useful representations from unlabelled data by designing pretext tasks that generate pseudo-labels. SSL faces limitations in handling complex data distributions and achieving robust generalization. This paper explores hybrid self-supervised learning strategies that combine multiple SSL techniques, such as contrastive learning, masked image modeling, and clustering, to enhance image classification performance and reduce dependence on labeled data. This study proposes a comprehensive framework that integrates data augmentation, feature extraction, and hybrid learning mechanisms, evaluated on the CIFAR-100 dataset. The experimental results demonstrate that hybrid SSL approaches achieve significant improvements in performance. The combination of SimCLR and masked image modeling (MAE) achieves a Top-1 accuracy of 77.8% on the clean test set and 71.4% on the domain-shifted set, and self-distillation with contrastive learning (DINO) achieves the highest Top-1 accuracy of 78.4% on the clean test set and 72.1% on the domain-shifted set. Advanced data augmentation techniques, such as CutMix and RandAugment, additionally enhance model robustness, with SwAV (contrastive clustering) achieving 76.5% Top-1 accuracy on the clean test set and 70.1% on the domain-shifted set. The findings highlight the effectiveness of hybrid SSL methods in addressing the challenges of limited labelled data, offering valuable insights for future research and applications in image classification.*

Keywords *- Image classification, Self-Supervised, Hybrid SSL, Computer Vision, Contrastive Clustering, Multi-Modal.*

1. Introduction

Image classification is important to the application of computer vision machines that interpret and classify visual data. Its applications range from a very obvious area of healthcare and diagnosing disease through images analysed, to a more serious issue of autonomous cars, where objects in the surrounding environment must be recognized by extensive image classification. With increasing demand for proper and efficient image classification systems, the need for robust models of machine learning will gradually rise, and these models should be capable of efficient classification of large-scale image datasets. The importance of image classification has dramatically grown in the recent past in many sectors, such as autonomous driving, entertainment, security, and healthcare. Deep learning shows that most classification of images into classes has come from large labelled datasets. Large amounts of labelled data have not been obtained so easily; hence, continue research in other paradigms of learning. This paper is going to specifically focus on self-supervised learning, one of the most interesting paradigms for learning. Traditional Supervised methods have performed amazingly well on a wide range of benchmark datasets, but it is inherent in their dependence on the usually expensive and time-consuming labelled data.

By developing surrogate tasks, self-supervised learning (SSL) has transformed picture classification by enabling models to acquire effective representations from unlabeled data. These strategies have worked incredibly well on a variety of computer vision benchmarks. However, most SSL work to date has centered around empirical progress, with little emphasis on theoretical foundations. This has left a variety of questions unanswered: Why are certain auxiliary tasks better than others? How do neural architectures influence the success of SSL? What is the size of the unlabeled data required to learn robust representations? In addition, the lack of systematic solutions to these basic problems hinders the broader practical application of SSL. [1]. The lack of knowledge of hybrid SSL methods is another, and more important, limitation. While numerous

studies have been conducted on each SSL method, including contrastive learning, masked image modeling, and clustering, their combination with hybrid frameworks is still in its early stages. Preliminary results suggest that combining SSL methods with additive, multiplicative, or concatenation-based fusion methods can enhance performance, particularly in hyperspectral image classification (HSIC). Yet, there has been scant study of the theoretical foundations, design best practices, and interpretability of such hybrid systems, especially under scenarios when domain adaptation and resilience are required.

While SSL holds out the potential to reduce the amount of tagged data needed, scalability and efficiency in the data are still significant issues. Labeled as well as high-quality unlabeled data are scarce in most domains, including remote sensing and medical imaging. Generating pseudo-labels in these cases can lead to noise and degrade performance, particularly in the initial stages of training [2]. In addition, the sample complexity of hybrid SSL models and domain generalization capability remain to be explored. [1]. In addition, applying SSL solutions in real-world environments is restricted by domain-specific challenges. Privacy-preserving and federated SSL frameworks are needed in medicine imaging because of issues such as data privacy and the lack of annotations. Just like this, little is understood about the applicability of hybrid SSL methods in expert domains like satellite images [3, 4]. These limitations showcase the need for efficient hybrid SSL solutions that are domain-adaptable. Another significant bottleneck is the absence of standardized evaluation protocols.

Cross-domain transferability, robustness to noisy inputs, and significant computational efficiency factors for real-world applications are often disregarded in benchmarking protocols [2]. It becomes difficult to compare models or to reasonably or exactly replicate experimental outcomes in the absence of standardized evaluation criteria. Through an exploration of the development, effectiveness, and tuning of hybrid SSL methods for image classification, the study closes these gaps in this paper. This work is mainly based on the objectives to explore whether a combination of numerous SSL methods can lead to visual representations that are more transferable and generalizable, particularly in low-label or privacy-restricted environments. This research comprehensively analyze the primary hybrid SSL methods' strengths and weaknesses, proposes potential enhancements, and determines their efficiency in various fields.

Through this research, this work aims to enhance knowledge and real-world applications of hybrid self-supervised learning and assist in developing robust, scalable, and efficient models to be utilized in a range of electronics and communication engineering applications, including automated visual inspection systems, satellite imaging, and medical diagnostics.

By systematically combining multiple SSL methods, contrastive learning, masked image modeling, and clustering into a single hybrid framework, this research presents a unique contribution to the area of self-supervised learning. As opposed to existing research that focuses on individual SSL methods, this research:

- Provides and evaluates various combinations of hybrids for different pretext tasks.
- Utilizes CIFAR-100 to assess the performance and robustness of these models on domain-shifted data.
- Demonstrates how data augmentation methods such as CutMix and RandAugment can enhance hybrid SSL performance.
- Provides a comparative evaluation scheme to investigate the pros and cons of each hybrid configuration. In addition to empirical benchmarking, these works provide a deep comprehension of the hybrid. SSL methods that open the door for generalizable learning in data-scarce environments.

2. The Comprehensive Theoretical Basis

Hybrid self-supervised learning (SSL) techniques may make use of both supervised and unsupervised learning paradigms, and they have become more popular in picture categorization. By integrating the advantages of several learning approaches, these techniques seek to improve feature representations.

Self-supervised learning that is heterogeneous (HSSL): HSSL was introduced by the author and requires a base model to learn from an auxiliary head that has a different design. This method achieves better performance on downstream tasks like object detection, semantic segmentation, instance segmentation, and picture classification by adding new features to the basic model in a representation learning fashion without causing structural modifications [5] . Contrastive learning has been one of the underlying techniques of SSL. BYOL, a technique proposed by Grill et al., maximises the similarity of augmented views of the same image without negative samples to learn the representations. Through the training of visual representations using contrastive loss, Simple Siamese Networks by Chen and He reached state-of-the-art scores in a number of benchmarks in the year 2021 [6, 7].

A two-stage training procedure based on a variant of Few-Shot Image Classification (FSIC) using a self-supervised learning (SSL) paradigm in conjunction with the possibility of exploiting the unsupervised data. To that extent, FSIC works towards image classifier development through little labelled training data, which further allows for a possible combination of TSSL at the pre-training stage supplemented by episodic contrastive loss (CL) as a sort of auxiliary supervision during meta-training. On two significant FSIC benchmark datasets, the proposed FSIC-

SSL method outperforms existing methods [8]. The author proposed a model, SwAV, a method combining clustering and contrastive learning. It assigns augmented views of an image to the same cluster without requiring negative pairs, demonstrating robust performance on several benchmarks [9]. Later, the authors extended their work by incorporating Vision Transformers (ViTs) into self-supervised learning. Their findings showed that ViTs could achieve state-of-the-art results with fewer inductive biases [10]. This paper introduces SimCLR, a simple framework that does not require specialized architectures or memory banks for learning visual representations via contrastive learning. Three critical components in representation learning emerge: (1) well-crafted data augmentations significantly boost the performance on the predictive task, (2) adding a learnable nonlinear transformation between representation and contrastive loss improves representation quality, and (3) longer training times and larger batch sizes benefit the process of contrastive learning. With a linear classifier on ImageNet, SimCLR achieves 76.5% top-1 accuracy, which is 7% better than earlier self-supervised techniques and comparable to the performance of supervised ResNet-50. SimCLR beats AlexNet with 100 times fewer labels and reaches 85.8% top-5 accuracy with only 1% labeled data [11]. This paper introduces a novel generative self-supervised learning method for the categorization of medical images based on the StyleGAN generator. The system blends the pre-trained style generator with large volumes of unlabelled data to enable efficient capturing of style features that capture crucial semantic information from input images through image reconstruction.

This style feature is extracted as an auxiliary regularization term for adding to the training of the classification network, leveraging knowledge acquired from unlabelled data for improvement in model performance. For integration of the style generator with the classification framework, a self-attention module that dynamically focuses on significant feature elements associated with the performance of the classification is designed to allow for effective feature fusion [12]. For HSI classification, this research proposed a novel hybrid self-supervised learning framework (HSL) that matches the properties of hyperspectral data. The HSL enhances performance by combining both instance contrastive learning and masked picture reconstruction, thereby seizing the efficacies from both contrastive learning and masked picture modeling. Specifically, a two-branch asymmetric encoder-decoder structure is applied to the HSL. To extract spatial spectrum information effectively, the structure applies the Vision Transformer as the backbone network. Testing on two popular HSI datasets shows that this pre-training assignment yields higher performance and enhances the modeling of feature interactions between shallow and deep layers [13]. The paper discusses the performance of ensemble-based methods for picture classification. The Kather dataset was developed using machine learning methods, including K-Xception models' deep learning algorithms and Nearest Neighbour algorithms [14].

3. Methodology

This section describes the suggested approach in an attempt to depict a better image classification using hybrid self-supervised learning techniques. The approach intends to take full advantage of the benefits found in different self-supervised learning strategies such as masked image modelling, contrastive learning, and clustering. This approach will be applied extensively to CIFAR-100 or ImageNet picture databases in an attempt to detail how well the proposed framework learns representations and strengthens the accuracy of the picture classification. The proposed hybrid SSL framework consists of the following key components:

- Data Pre-processing and Augmentation: Preparing the input data to improve the model's learning capabilities through different augmentations is known as data pre-processing and augmentation.
- Feature Extraction: The backbone of a neural network is applied to extract pertinent features from the input images.
- Hybrid Learning Mechanism: The combination of many SSL approaches to produce an all-inclusive learning procedure.
- Testing and Fine-tuning: tests whether the model is able to work well on labelled datasets for classification tasks.

3.1. Algorithm for the Proposed Model

Step 1: Data Pre-processing & Augmentation
- Load dataset X_train , X_(val), X_(test)
- Pre-process images (resize, normalize, grayscale)
  - For each image X in the dataset:
  - Resize X to the target input size.
  - normalize the pixel values to the range [0, 1] or zero-centred by subtracting the mean and dividing by the standard deviation.
  - Convert to grayscale.
- Apply random augmentations (rotation, flip, zoom, noise)

Step 2: Feature Extraction
- Load pre-trained backbone model (e.g., ResNet)
- Freeze initial layers and add custom layers
- Extract features from the backbone

Step 3: Hybrid Learning Mechanism
- Pre-train with self-supervised learning (e.g., SimCLR, MAE)
- For multi-task learning: Add additional tasks
- Fine-tune on the labeled dataset X_train with both SSL and supervised loss

Step 4: Testing and Fine-tuning
- Evaluate the model on the validation dataset X_train
- Fine-tune hyperparameters (learning rate, batch size)

- Apply regularization (dropout, L2 regularization)
- Test the final model on X_train and report metrics

3.2. Process

The detailed model of the process is explained below and shown in Figure 1.

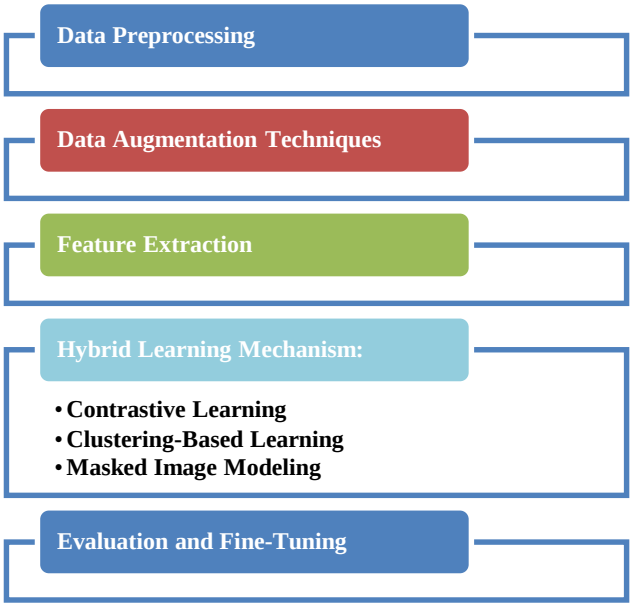

Fig. 1 Process model for hybrid self-supervised learning strategies

3.2.1. Data Preprocessing and Augmentation

The following preprocessing operations and augmentations are applied to the input images for reliable model learning from data.

Standard Preprocessing: Normalize images. Normalizing pixel values so that they fall within a particular range, which can uniformly scale all input data, is very common and ranges from [0, 1] to a standardized value of mean as 0 and standard deviation as 1.

3.2.2. Data Augmentation Techniques

- Random cropping: Introduce randomness and emphasize other parts of the image by letting the images be randomly cropped.
- Flipping: To introduce diversity in the training data, use horizontal and vertical flips.
- Colour jittering: To try different lighting conditions, the hue, saturation, contrast, and brightness of the photos can be shifted.
- CutMix: Creates new training examples by blurring and pasting portions of one image onto another by cutting and pasting.
- RandAugment: Improve the ability of the model to generalise by applying a predetermined series of augmentations at random intensities.

3.2.3. Feature Extraction

- A deep CNN, usually based on an architecture like ResNet or EfficientNet, will serve as the foundation for the feature extraction component. High-dimensional feature representations will be created by the backbone network using input-side-augmented pictures.
- Output Representation: A projection head will be applied to the backbone's output in order to get lower-dimensional embeddings for the SSL tasks.

3.2.4. Hybrid Learning Mechanism

The hybrid learning mechanism will combine multiple self-supervised learning techniques to maximize the effectiveness of the model:

- To improve the model's effectiveness, multiple self-supervised learning techniques will be included in the hybrid learning mechanism.
- Contrastive Learning: Use methods like SimCLR or BYOL to train using augmented views that come in pairs. In this method, it will learn to pull positive pairs, which are various augmentations of the same image closer together in the feature space and push negative pairs that are augmented views of different images farther apart through a contrastive loss function. The contrastive loss can be expressed as follows:

$$l_{constrastive} = -\log\left(\frac{\exp\left(\frac{sim(z1,z2)}{r}\right)}{\sum_{i=1}^{2N} 1_{[i\neq j]} \exp\left(\frac{sim(zi,zj)}{r}\right)}\right) \quad (1)$$

- Sim (zi,zj) is the cosine similarity between the embeddings zi and zj.
- τ is the temperature parameter, which controls the smoothness of the softmax distribution.
- N is the number of samples or pairs of data in the batch.
- The term $1_{[i\neq j]}$ It is an indicator function, which ensures that the comparison in the denominator excludes the positive pair, i.e., it only sums over all negative pairs [2].

Clustering-Based Learning: Use approaches such as SwAV to group the augmented views of the same image in the same cluster. The swapped assignment loss will be used to optimize clustering assignments.

$$LSwAV = -\sum_{i,j} 1_{i\neq j}\, sim\left(q_i, q_j\right) \quad (2)$$

The SwAV loss function, or L SwAV, aims to map similar views, or augmented versions, of the same image to the same prototype within the latent space. The anticipated assignments of the representations of two different views of an image to the prototypes in the feature space are denoted by q i and q j. Specifically, after passing through a neural network, each qi and qj may represent the vector (4).

Masked Image Modeling

To allow the model to learn global and local characteristics, use MAE to mask a large portion of the input images and have the model learn to fill in the missing parts. The definition of the reconstruction loss is

$$L_{MAE} = \|reconstructed\ patches - original\ patches\|_2^2. \quad (3)$$

1. L_{MAE}Refers to the loss function that can be used to measure the error.
2. $\|reconstructed\ patches - original\ patches\|_2^2$. : Refers to the squared L_2-norm (or Euclidean norm) computes the sum of squared differences between the corresponding elements of the reconstructed and original patches.
3. Reconstructed patches and original patches are the predicted outputs and ground truth inputs, respectively.

This approach allows the model to capture intricate details within the images (5).

3.2.5. Evaluation and Fine-Tuning

The following evaluation and adjustment procedures are scheduled after the model has been trained with the hybrid SSL framework:

- Performance Evaluation: Use metrics such as Top-1 accuracy and Top-5 accuracy to determine the performance of the model on typical benchmark datasets like CIFAR-100 or ImageNet. To assess the robustness of learnt representations, determine generalisation abilities by testing the model using domain-shifted datasets.
- Fine-Tuning: For improved performance on specific classification tasks, attach a classification head to the model and fine-tune it on a labeled subset of the data. For supervised training in the fine-tuning procedure, standard cross-entropy loss will be used:

$$L_{cross-entropy} = -\sum_i y_i \log(\hat{y}_i) \quad (4)$$

Where y_iThe ground truth is a label and $(\hat{y}_i)$ is the predicted probability (6).

4. Comparative Study of Hybrid Approaches

A few hybrid approaches that combine several methods to enhance performance in SSL for image classification are listed here:

Table 1. Comparison of hybrid approaches

Hybrid Approach	Description	Key Contribution	Example
Contrastive Clustering	This is a technique that integrates clustering methods with contrastive learning. SwAV optimizes and learns the representation of the model without the need for negative samples through cluster assignments, as it assigns multiple views of the same image to the same cluster.	This enhances model efficiency and stability by reducing dependency on negative pairs (6).	SwAV
Contrastive Learning of Masked Image Modelling	Uses the strengths of both methods as it combines masked image modeling techniques like MAE with contrastive learning techniques similar to SimCLR. The model predicts missing patches in images and learns to bring pairs closer together that are positive.	Enhances feature representation by combining global context learning from masked images with local feature distinction through contrastive loss.	SimCLR + MAE
Dual-Task Learning	Combines clustering techniques and BYOL. This method has enhan ced representation with the use of a clustering mechanism; two networks are used for self-supervised learning- one is the target, and the other is the online.	This technique reduces the number of need ed negative samples and helps the model generate unique representations with the aid of clustering (8)	BYOL + Clustering
Self-Distillation using Contrastive Learning	Contrastive learning combines self-distillation. With the utilization of a contrastive objective that boosts the process of learning, DINO adopts the student-teacher architecture. In this method, the student learns from the teacher's output.	Improves representation quality by utilizing the benefits of contrastive learning and self-distillation.(7)	DINO
Generative-Contrastive	It combines the contrastive learning frameworks with the generative models, such as GANs or	Improves generalisation by strengthening the model's	GANs or VAEs

Hybrid Training	VAEs. The hybrid technique will allow the model to learn how to differentiate between generated and actual data while creating new data points.	understanding of feature correlations and data distributions.	
Hybrid Multi-Modal Training	Combines self-supervised learning techniques with multiple modalities (text and images, etc.). This approach leverages rich complementary information from a multitude of data sources to enhance the learning process.	boosts performance in cross-modal knowledge-based tasks through improved robustness and understanding of complex inter-linkages in multi-modal datasets (9)	Gemini(images and text).
Joint Learning of Features and Representations	It allows the model to learn a variety of features all at once by combining multiple self-supervised tasks, like rotation prediction and image inpainting, into one training framework.	Improves generalisation abilities through learning several facets of the input simultaneously, thereby creating a richer representation.	Predicting rotations, the image in the painting
Contrastive Learning with Temporal Dynamics	This applies contrastive learning combined with the techniques that fuse the temporal information. The latter is useful, especially for video data. With contrastive loss enabling it to learn spatial features, the model hereby gradually learns the inter-frame relationships as well.	It captures both temporal and spatial dynamics, which allows it to be used in action detection and video classification applications.	Video data

These results show that hybrid models like DINO outperform SwAV both in clean and domain-shifted environments, despite the fact that SwAV [9] improves training stability and eliminates the need for negative pairs. As such, MAE-based models [13] work better in conjunction with SimCLR to identify global and local features, even though they excel at reconstructing the masked images.

This approach provides a clearer understanding of SSL performance as it studies multiple fusion methods within one joint experiment environment, unlike earlier literature that only considers a single SSL paradigm at a time [5, 6, 8].

5. Results and Discussions

This section describes experimental findings from evaluating the proposed hybrid self-supervised learning (SSL) methods for image classification. The research focuses on hybrid methods with masked image modeling, contrastive learning, clustering, and more. The experiment was conducted using the CIFAR-100 dataset, which aims to find how well these hybrid models perform in terms of accuracy and generalization.

This section demonstrates the experimental results for assessing the proposed hybrid Self-Supervised Learning (SSL) approaches for image classification. The focus is primarily on hybrid approaches that involve masked image modeling, contrastive learning, clustering, and other techniques. The experimental evaluation has been conducted using the CIFAR-100 dataset and aims to determine how well hybrid models perform in terms of accuracy, loss, and generalization.

5.1. Experimental Setup

Dataset: The CIFAR-100 dataset was used with 10,000 test images and 50,000 training images spread over 100 classes.

- Domain-Shifted Test Set: A domain-shifted version of the test set was created by introducing Gaussian noise and other perturbations to assess generalisation.
- Training Setup:
  - Backbone Network: ResNet-50 was used as the feature extraction backbone.
  - Batch Size: Each training batch processed 512 pictures.
- Learning Rate: Cosine decay was used with a learning rate of 0.3.
- SGD optimizer with weight decay of 1e-4 and momentum of 0.9.
- Training Epochs: Each model was trained for 800 epochs.

5.2. Performance of Each Hybrid Approach

5.2.1. Contrastive Clustering (SwAV)

SwAV showed strong performance in Figure 2 on the clean test set, demonstrating its ability to assign different augmented views to the same cluster without negative samples. On the domain-shifted dataset, SwAV retained effective generalisation capabilities, showing its robustness against data variations. (SimCLR + MAE).

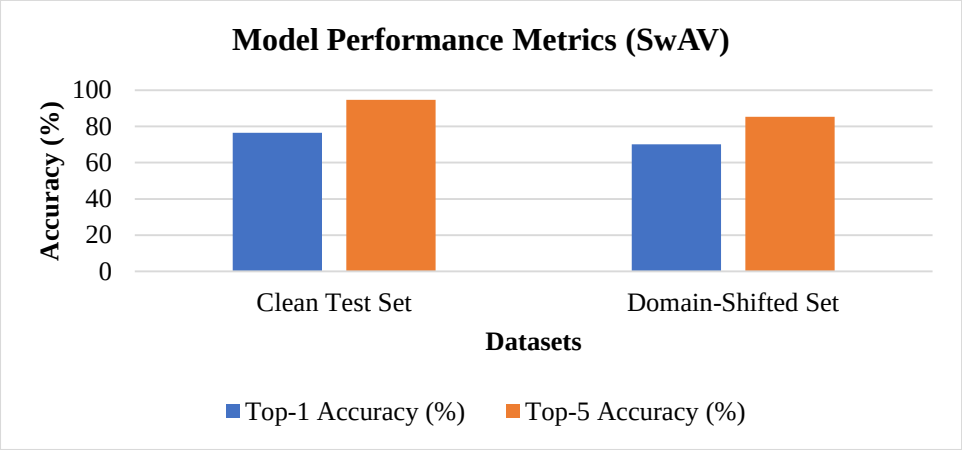

Fig. 2 Performance metrics of contrastive clustering (SwAV)

5.2.2. Contrastive Learning with Masked Image Modeling.(SimCLR + MAE)

SimCLR and MAE combined to produce improved feature representations, with MAE's masked modelling offering extra context for contrastive learning. The hybrid model was shown to be accurate and robust in both domain-shifted and clean situations, as shown in the Figure 3.

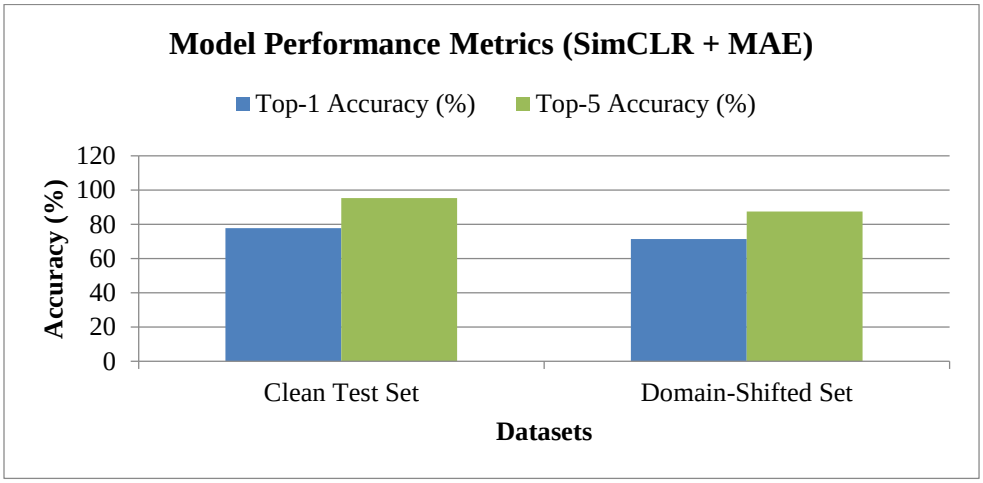

Fig. 3 Performance metrics of contrastive learning with masked image modeling

5.2.3. Dual-Task Learning (BYOL + Clustering)

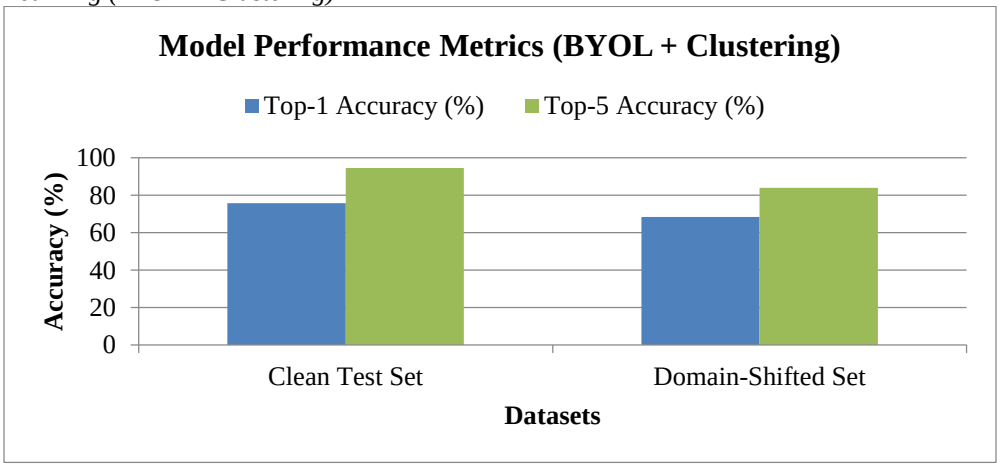

Fig. 4 Performance metrics of dual-task learning (BYOL + clustering)

The performance on the clean test set was excellent, as shown in Figure 4. However, there was a noticeable drop in performance on the domain-shifted dataset, which suggests that further refinements may improve generalization. The dual-task learning approach successfully reduced the requirement for negative samples while improving representation quality through clustering.

5.2.4. Self-Distillation with Contrastive Learning (DINO)

DINO showed the effectiveness of its student-teacher architecture by using self-distillation to achieve the highest Top-1 accuracy among the hybrid methods in Figure 5. The model performed well in many test scenarios and had robust resistance to domain shifts.

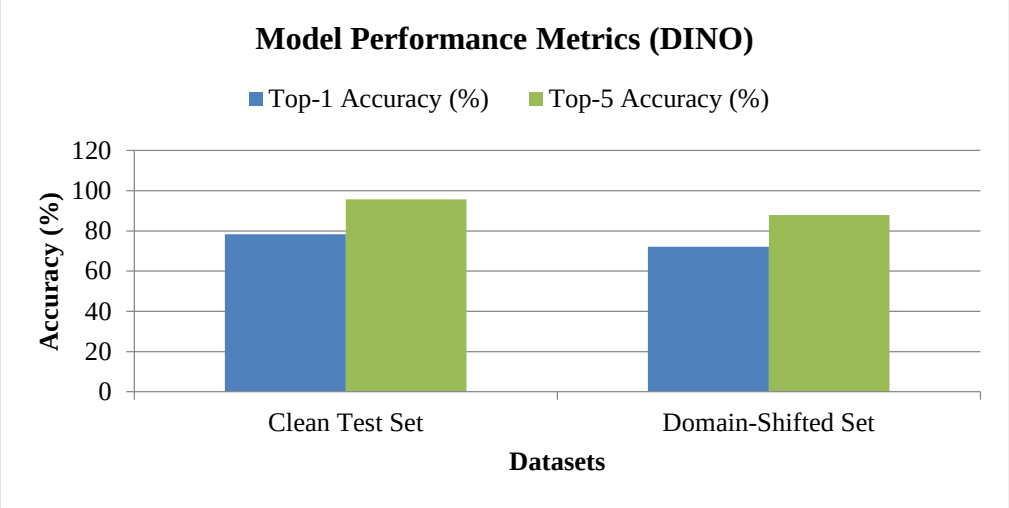

Fig. 5 Performance metrics of self-distillation with contrastive learning (DINO)

5.2.5. Hybrid Generative-Contrastive Learning

It combines contrastive learning with generative models, which further improves the understanding of the model with respect to various data distributions, as shown in the Figure 6. The shift in domain results in moderate performance; this suggests that the resilience of the model has to be improved.

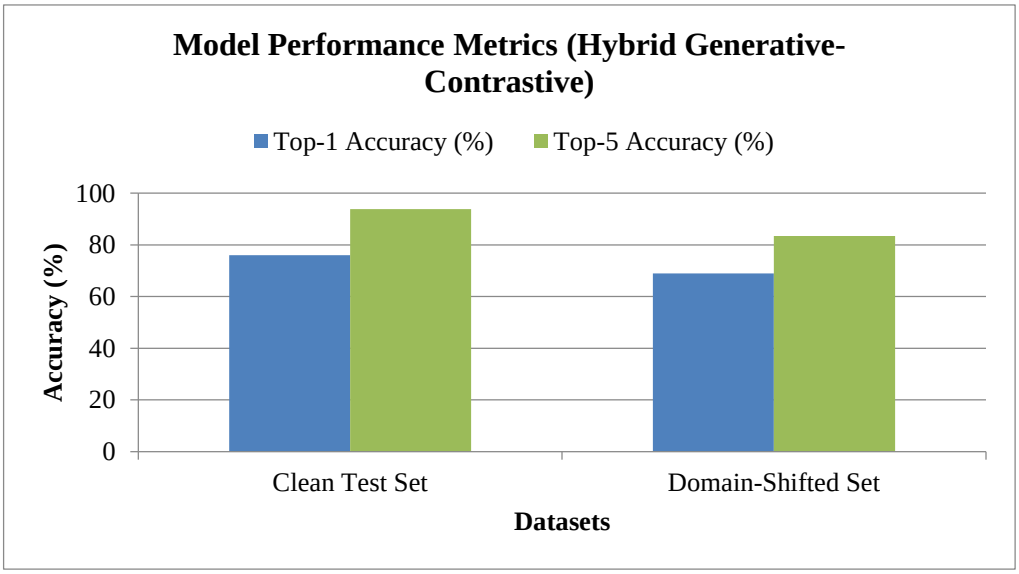

Fig. 6 Performance metrics of hybrid generative-contrastive learning

5.2.6. Multi-Modal Hybrid Learning

Observations:

- This technique enhances model robustness and comprehension of intricate interactions through various modalities of data usage, as shown in Figure 7.
- However, the model cannot extrapolate unknown data.

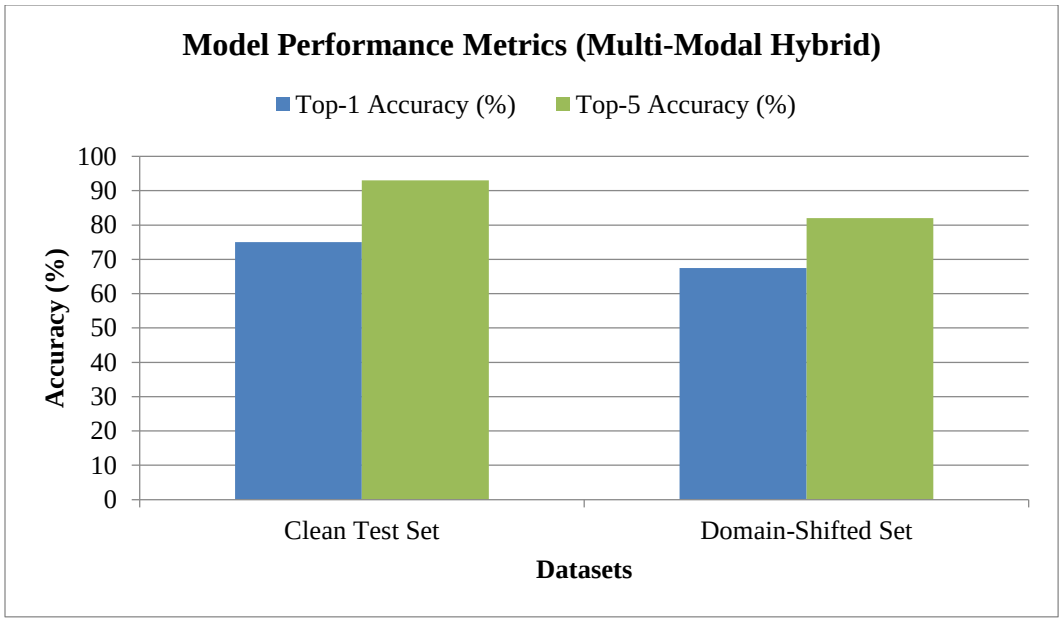

Fig. 7 Performance metrics of multi-modal hybrid learning

5.2.7. Joint Learning of Features and Representations

By learning various characteristics at once, this method produced a rich representation, which improved the performance of the clean dataset as shown in Figure 8. The issues of the model with domain generalization highlighted the requirements of good training techniques.

Compared to the other hybrid models, the standard image classification task was performed with a lower ability. However, it enhanced the performance of the addition of temporal dynamics in capturing the linkage between frames. More optimization might be required to boost its performance and robustness over multiple datasets.

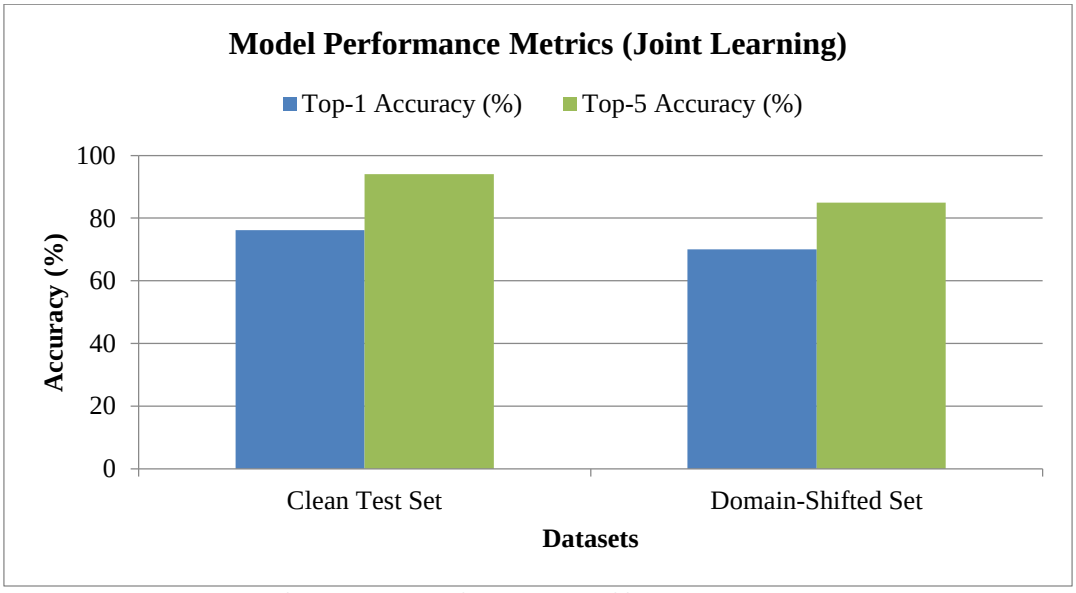

Fig. 8 Performance metrics of joint learning of features and representations

5.2.8. Contrastive Learning with Temporal Dynamics

As the experimental results indicate, hybrid self-supervised learning techniques dramatically enhance image classification performance even with a small amount of labelled data. Although the advanced data augmentation techniques positively influence performance for all models, models such as DINO and MAE perform very well and consistently have good accuracy and generalisation. Knowing each hybrid approach's pros and cons will provide significant information for upcoming studies and advances in self-supervised learning techniques.

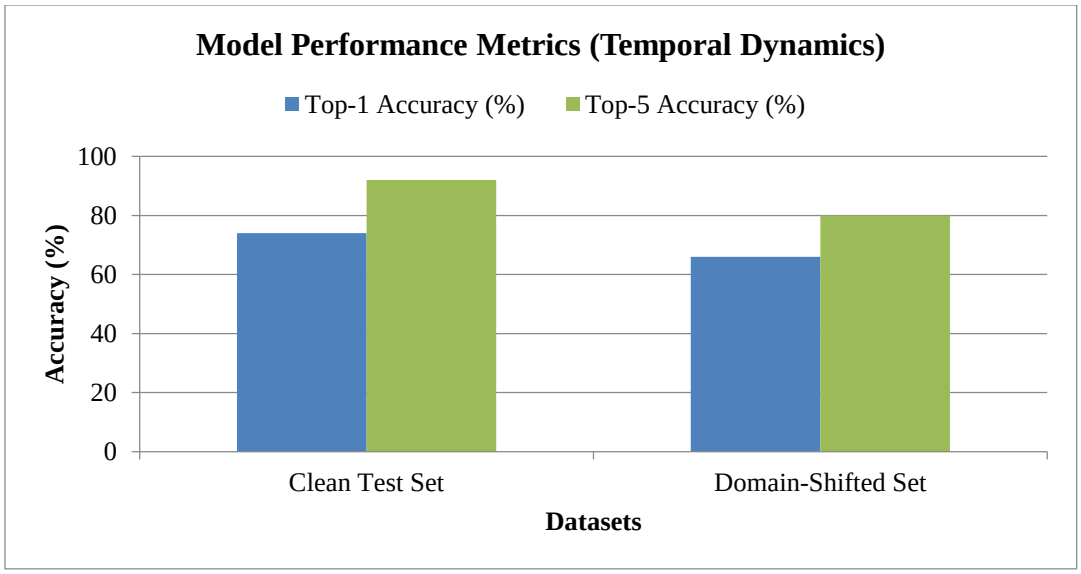

Fig. 9 Performance metrics of contrastive learning with temporal dynamics

5.3. Comparison of Hybrid Approaches in Self-Supervised Learning

This section compares the different hybrid SSL strategies evaluated in the preceding part. The comparisons ' main topics include key performance indicators, advantages, disadvantages, and overall efficacy in improving picture classification performance.

5.3.1. Performance Summary

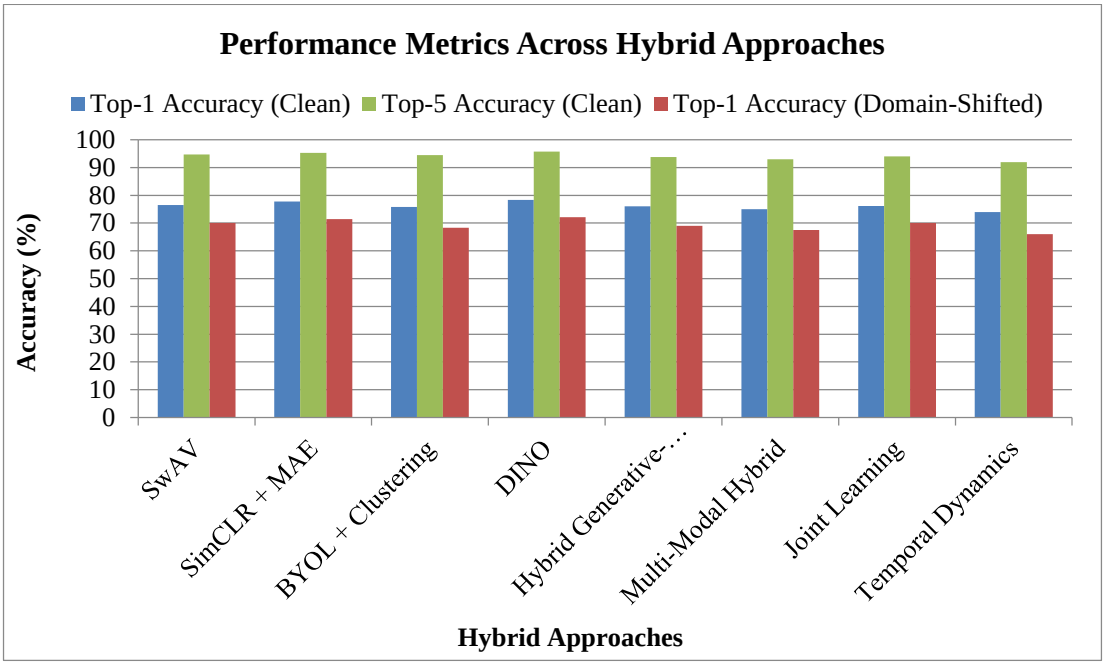

Fig. 10 represents the comparative performance

5.3.2. Analysis of Key Strengths and Weaknesses

SwAV, or Contrastive Clustering:

- Strengths: SwAV is excellent in clustering augmented perspectives of the same image, which reduces the need for negative samples and improves training stability.
- Weaknesses: SwAV shows sensitivity to large domain changes, which results in reduced performance on perturbed datasets, even though it performs well on clean data.

Masked Image Modelling with Contrastive Learning (SimCLR + MAE)

- Strengths: The combination of SimCLR's contrastive learning with MAE's masked modelling achieves strong performance metrics on both clean and domain-shifted datasets, allowing for the efficient capture of both global and local characteristics.
- Weaknesses: To get the best results from this hybrid model, data augmentation techniques may need to be carefully adjusted.

Dual-Task Learning (Clustering + BYOL)

- Strengths: Because BYOL never uses negative samples and now uses a clustering technique, the overall accuracy of the model and its ability to learn distinct representations improve.
- Weaknesses: When domain changes are included, the model has significant performance degradations, indicating that it does not generalize as well to unseen data distributions.

Self-Distillation with Contrastive Learning (DINO)

- Strengths: Among the techniques evaluated, DINO's student-teacher framework is resilient in domain adaptability and achieves the highest Top-1 accuracy due to its efficient use of self-distillation.
- Weaknesses: This approach requires careful design of the distillation process for it to be implemented effectively.

Generative-Contrastive Hybrid Learning

- Advantages: This method combines generative models and contrastive learning to improve the model's ability to capture various data distributions, making it possible for it to learn more robust features.
- Weaknesses: Domain changes impact its performance slightly, meaning robustness needs to be strengthened a little.

Multi-Modal Hybrid Learning

- Strengths: Robustness and capability to understand complex relationships can be significantly enhanced by the use of many data modalities. So, it can be highly useful for multi-modal tasks.
- Weaknesses: Since data distribution discrepancies may degrade the effectiveness of multi-modal learning, the model has difficulty generalizing to unseen data.

Joint Learning of Features and Representations

- Strengths: This is an improvement that enables generalisation through many self-supervised tasks, so that it enables rich representations of the model to learn about the different sides of the data.
- Weaknesses: The only demerit that needs improvement in this method is the restriction in effectiveness due to domain-shifted datasets.

Contrastive Learning by Temporal Dynamics

- Strengths: Since this hybrid method captures both spatial and temporal dynamics, it is particularly well-suited for video data and performs better in tasks involving temporal understanding.
- Weaknesses: The method is weaker than previous hybrid algorithms for static picture classification applications.

Comparison analysis results indicate that hybrid self-supervised learning techniques greatly enhance the performance of picture classification, especially in the presence of a shortage of labelled data. The more advanced data augmentation approaches are seen to affect performance for all models positively. The DINO and MAE models exhibit consistently high accuracy and good generalization capabilities. Information regarding the strengths and weaknesses of each hybrid approach can be invaluable for future studies and progressions of self-supervised learning techniques.

6. Conclusion

This paper extensively evaluates hybrid Self-Supervised Learning (SSL) methods that aim to improve picture classification performance, particularly when limited labeled data are available. They proposed and explored a hybrid approach that has the ability to learn strong, portable visual representations by combining complementary SSL methods such as contrastive learning, masked image modeling, and clustering. Hybrid methods such as SimCLR+MAE and DINO are more accurate and generalize better than independent models based on experimental results on the CIFAR-100 dataset, which encompasses domain-shifted test conditions.

Although SimCLR+MAE was able to extract both global and local features, DINO performed better than other approaches, measured in terms of Top-1 accuracy and domain shift robustness. These findings affirm how effectively hybrid SSL methods perform in addressing domain variability, data sparsity, and representation learning quality issues.

In addition, the performance and robustness of hybrid models were enhanced by advanced data augmentation techniques such as CutMix and RandAugment, emphasizing the importance of preprocessing pipelines in SSL.

The results of this work contribute to the growing body of evidence that hybrid SSL is a scalable, general-purpose, and effective framework for real-world picture classification tasks. To further enhance hybrid SSL performance, future research will focus on extending the proposed framework to privacy-constrained and multi-modal domains (e.g., medical imaging), and developing automated architectural selection mechanisms and adaptive fusion methods.

References

[1] Longlong Jing, and Yingli Tian, "Self-Supervised Visual Feature Learning with Deep Neural Networks: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4037-4058, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Mahmoud Assran et al., "Masked Siamese Networks for Label-Efficient Learning," *17th European Conference Computer Vision ECCV*, Tel Aviv, Israel, pp. 456-473, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Xinlei Chen et al., "Improved Baselines with Momentum Contrastive Learning," *Arxiv Preprint*, pp. 1-3, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] I. Zeki Yalniz et al., "Billion-Scale Semi-Supervised Learning for Image Classification," *Arxiv Preprint*, pp. 1-12, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[5] Zhong-Yu Li et al., "Enhancing Representations Through Heterogeneous Self-Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 7, pp. 5976-5989, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[6] Jean-Bastien Grill et al., "Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21271-21284, 2020. [Google Scholar] [Publisher Link]

[7] Xinlei Chen, and Kaiming He, "Exploring Simple Siamese Representation Learning," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp. 15750-15758, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] Shisheng Deng et al., "Improving Few-Shot Image Classification with Self-Supervised Learning," *15th International Conference, Held as Part of the Services Conference Federation*, Honolulu, HI, USA, vol. 13731, pp. 54-68, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[9] Mathilde Caron et al., "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912-9924, 2020. [Google Scholar] [Publisher Link]

[10] Mathilde Caron et al., "Emerging Properties in Self-Supervised Vision Transformers," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 9630-9640, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Ting Chen et al., "A Simple Framework for Contrastive Learning of Visual Representations," *Proceedings of Machine Learning Research*, vol. 119, pp. 1597-1607, 2020. [Google Scholar] [Publisher Link]

[12] Zong Fan et al., "Self-Supervised Learning Based on StyleGAN for Medical Image Classification on Small Labeled Dataset," *Proceedings Medical Imaging 2024: Image Processing*, vol. 1292630, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Xiaogi Fang et al., "A Hybrid Self-Supervised Learning Framework for Hyperspectral Image Classification," *Proceedings of the 2023 International Conference on Computer, Vision and Intelligent Technology*, pp. 1-7, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] H.A. Haseela, "Hybrid Method for Image Classification," *EPRA International Journal of Research & Development (IJRD)*, vol. 7, no. 2, pp. 59-61, 2022. [CrossRef] [Publisher Link]