SCIENTIFIC INNOVATION RESEARCH GROUP
Journal of Smart Algorithms and Applications JSAA
INTERNATIONAL JOURNAL
SIRG

Journal of Smart Algorithms and Applications
JSAA
3070-4189/© 2026 JSAA. All Rights Reserved.
Journal Homepage
https://pub.scientificirg.com/index.php/JSAA

A Unified Hybrid Self-Supervised Architecture Integrating Contrastive and Non-Contrastive Learning for Multi-Level Visual Representations

Ali N. K Alhejoh[a], Ibrahim AbuNahleh[a1], Mohammed Hashem Almourish [a,b]

[a] Department of Computer Science, Isra University, Amman, Jordan, E-mails: ali.alhejoj@iu.edu.jo, ibrahim.abunahleh@iu.edu.jo
[b] Department of Information Technology, University of Taiz, Hubail Salman, Taiz, Yemen, E-mail: mohamed.almourish@iu.edu.jo

ABSTRACT - This paper proposes a unified hybrid self-supervised learning framework for fine-grained visual representation learning, evaluated on the STL-10 dataset. To address challenges arising from limited labeled data and high intra-class visual variability, the proposed approach integrates complementary contrastive and non-contrastive paradigms BYOL, SimCLR, MoCo v3, and DINO within a single architecture. The framework employs a shared backbone network with dedicated projection and prediction heads, alongside a momentum-updated target network using exponential moving average (EMA), enabling robust representation learning from unlabeled images. Diverse data augmentations, including random resized cropping, horizontal flipping, color jittering, and grayscale conversion, are used to generate multiple correlated views that jointly support hybrid training objectives. Experimental results show stable and consistent convergence, with an average BYOL loss reaching −0.88. Downstream evaluation demonstrates the effectiveness of the learned representations, achieving a test accuracy of 87.20%, a recall of 87.20%, and an F1-score of 87.19%. Additionally, the framework attains a mean Average Precision (mAP) of 93.70%, indicating strong discriminative capability and transferability. These results suggest that hybrid self-supervision can effectively exploit the complementary strengths of contrastive and non-contrastive learning, leading to improved representation quality and faster convergence compared to individual self-supervised methods.

PAPER INFORMATION

HISTORY
Received: 21 July 2025
Revised:15 November 2025
Accepted: 27 January 2026
Online: 8 February 2026
MSC
62K05
62K15

KEYWORDS
Contrastive learning
Representation learning
Non-contrastive learning
Unlabeled image datasets
c

Department of Computer Science, Isra University, Amman, Jordan, E-mail: ibrahim.abunahleh@iu.edu.jo

1. INTRODUCTION

The emergence of Self-Supervised Learning (SSL) has fundamentally transformed visual representation learning in computer vision [1], [11]. By leveraging the inherent structure of large-scale unlabeled datasets, SSL methods enable the extraction of high-level semantic representations without the need for extensive manual annotation[22]. Modern SSL frameworks typically project images into a latent embedding space, where representations are optimized to remain invariant across augmented views while preserving instance-level discrimination. This paradigm has proven essential for building scalable models that generalize effectively across a wide range of downstream visual tasks. Despite these advances, a critical limitation arises from the pursuit of strong invariance. Conventional approaches, particularly those based on the InfoNCE loss, are designed to minimize discrepancies between different augmented views of the same image. While this objective successfully captures global semantic concepts, it frequently suppresses fine-grained spatial information. This phenomenon, referred to as representational deficiency, occurs when the learning process prioritizes semantic alignment at the expense of local structure, thereby discarding geometric and positional cues that are crucial for tasks extending beyond coarse classification. As a result, the learned encoder becomes increasingly insensitive to spatial variations, limiting its effectiveness in applications that require localization, correspondence, or structural reasoning [6].

Motivated by this challenge, this work examines the balance between semantic consistency and spatial awareness in self-supervised representation learning. Rather than replacing existing methodologies, we propose a Hybrid Multi-Branch Architecture that integrates complementary SSL paradigms within a unified framework. Specifically, our approach combines contrastive learning methods such as SimCLR and MoCo v3, non-contrastive approaches including BYOL and VICReg, and patch-level self-supervision techniques such as DINO and iBOT. Through a jointly optimized weighted objective, the proposed framework captures both global semantic robustness and local structural detail. This integration enables the learning of representations that are semantically expressive while remaining sensitive to the geometric properties of the input data [3], [4].

This work makes four primary contributions to self-supervised visual representation learning. First, we introduce a unified hybrid architecture that coherently integrates contrastive, non-contrastive, and patch-level self-supervised learning paradigms within a single multi-branch framework, enabling multi-level feature extraction without architectural redundancy [10]. Second, we provide a formal mathematical characterization of representational deficiency by explicitly modeling the trade-off between semantic invariance and spatial information preservation, thereby offering theoretical insight into the limitations of strongly invariant objectives [24]. Third, we propose a principled strategy for multi-level feature fusion that combines global image-level representations with local patch-level embeddings, ensuring that the learned features retain both high-level semantic meaning and fine-grained structural information [2], [8]. Finally, we empirically demonstrate on the STL-10 benchmark that the proposed hybrid framework achieves improved convergence stability and produces higher-quality representations compared to single-objective SSL baselines [12], [14].

The remainder of this paper is organized as follows. **Section 2** reviews related work in contrastive and non-contrastive self-supervised learning, with particular attention to recent advances in patch-level supervision. **Section 3** presents the proposed hybrid model architecture, including the integration of a Vision Transformer backbone. **Section 4** describes the STL-10 dataset and the data preprocessing and augmentation strategies employed in training. **Section 5** establishes the mathematical formulation of self-supervised learning objectives and analyzes the limitations of invariance-based methods. **Section 6** details the proposed solution by outlining the loss functions associated with each learning branch. **Section 7** reports and analyzes the experimental results, including convergence behavior and representation quality evaluation. Finally, **Section 8** concludes the paper and discusses potential directions for future research in hybrid self-supervised learning frameworks.

2. LITERATURE REVIEW

TULIP Contrastive Image-Text Learning with Richer Vision Understanding. Self-supervised learning approaches such as SimCLR and MoCo learn visual representations by contrasting multiple views of the same image, primarily capturing global semantics. However, these methods often fail to model fine-grained visual details. Recent works, including DINOv2 and iBOT, address this limitation through patch-level self-supervision, enabling better spatial and local feature understanding.

Multimodal methods such as TULIP further enhance representation learning by aligning visual and textual features using contrastive learning. The integration of contrastive, non-contrastive, and patch-level learning results in more robust and comprehensive visual representations [1]. Self-Supervised Learning of Pretext-Invariant Representations. Some methods that teach computers to understand images by guessing how they might look after changes don't always work well because the changes can affect how well the computer understands the real meaning of the image. PIRL solves this by making sure the computer learns to see the same thing in both the original and changed images, which helps it understand better and perform better when applying what it learned. However, PIRL mainly focuses on the whole image and doesn't pay much attention to smaller details. Newer methods combine learning about the whole image with learning about small parts of it, allowing the computer to understand images more deeply without needing to guess what the image was meant to do [2]. DINO: Emerging Properties in Self-Supervised Vision Transformers DINO is a self-supervised method based on knowledge distillation without labels, using a teacher-student framework. It encourages consistency between global and local image views without relying on negative samples. DINO enables Vision Transformers to learn meaningful semantic representations. The learned features naturally capture object boundaries and spatial structures. This makes DINO effective for fine-grained and dense vision tasks such as segmentation [3]. iBOT: Image BERT Pre-Training with Online Tokenizer iBOT extends self-supervised learning to patch-level representations using a masked image modeling strategy. It enforces consistency between visible and masked patches through a teacher–student architecture. Unlike global-only methods, iBOT learns both image-level and token-level semantics. This design improves spatial understanding and local feature representation. iBOT achieves strong performance on dense prediction tasks.[4].

ICReg: Variance-Invariance-Covariance RegularizationVICReg is a non-contrastive self-supervised learning method that avoids negative samples entirely. It learns representations by enforcing invariance, variance, and decorrelation constraints. This method stabilizes training while preserving informative representations. VICReg primarily captures global semantics but can be extended to hybrid frameworks. It achieves competitive performance without collapse.[5]

BYOL: Bootstrap Your Own Latent BYOL is a type of self-supervised learning that doesn't use negative examples to learn visual features. It works with a teacher and student setup, where the student tries to predict what the teacher sees, even when the images are changed in different ways. BYOL shows that good representation learning can happen without using contrastive loss. The features it learns are good at capturing overall meaning, but don't focus much on smaller parts of images. BYOL has inspired many mixed models that combine both global and local learning [6].

SimCLR: A Simple Framework for Contrastive Learning of Visual Representations SimCLR is a type of self-supervised learning method that helps computers understand images by making them agree on different ways of looking at the same picture. It focuses on learning overall meaning from the image. But SimCLR needs big groups of data and a lot of examples that are not similar. It doesn’t pay much attention to small parts or how things are arranged in space. Later work improved on SimCLR by adding features that look at different levels or parts of the image [7].

MoCo v3: An Empirical Study of Training Self-Supervised Vision Transformers MoCo v3 applies contrastive learning to Vision Transformers by using a teacher-student setup that relies on momentum. It eliminates the need for memory banks but still follows contrastive learning goals. MoCo v3 develops strong overall representations but does not directly ensure detailed spatial consistency. This issue leads to the use of patch-level and non-contrastive learning methods. [8]

SwAV: Unsupervised Learning of Visual Features by Contrasting Cluster Assignments SwAV is a self-supervised learning method that uses clustering and doesn't compare pairs of images. It learns to represent images by guessing which cluster each image belongs to, using different views of the same image. SwAV is good at capturing meaningful information and works well even with smaller groups of images. However, it mainly focuses on understanding images. To better understand the details within images, more work is needed at the level of smaller parts of the image. [9]

MAE Masked Autoencoders Are Scalable Vision Learners. MAE is a type of self-supervised learning that doesn't use contrast. It works by hiding parts of an image and then trying to guess what's missing based on the rest of the image. This helps the model learn detailed features from local areas. MAE is good at understanding small details and how things are arranged in space. But it doesn't do a good job of connecting different parts of the image in a meaningful way. Using MAE with other methods like contrastive learning or distillation can make the model's learning more reliable [10].

BEiT: BERT Pre-Training of Image Transformers BEiT uses masked token prediction on images with discrete visual tokens. It learns detailed representations at the patch level through self-supervised training. BEiT focuses on understanding local meaning but does not ensure consistency across different image views created through augmentation. Mixed frameworks combine BEiT's masking approach with contrastive goals to improve representations at various levels. [11], as shown in **Table 1**

Table 1: Comparison of Self-Supervised Learning Methods Used in This Study

Ref	Methos	Dataset	Accuracy (Metric)	Contribution	Limitation
[1]	Robust Homography Estimation Network	Benchmark homography datasets	Homography error	improves robustness to noise and outliers in homographs estimation	Computationally expensive
[2]	Contrastive Representation Learning	ImageNet	Contrastive loss / Accuracy	Learns strong visual representations through instance discrimination	Performance depends on large batch sizes
[5]	VICReg	ImageNet	Linear evaluation accuracy	Prevents representation collapse without using negative samples	Requires careful balancing of loss components
[4]	iBOT (Image BERT Pre-training)	ImageNet	Top 1 accuracy	Introducing an online tokenizer for effective self-supervised learning	Sensitive to hyperparameter tuning
[3]	Self-Supervised Vision Transformer (DINO)	ImageNet	Linear probing accuracy	Demonstrates emergent semantic properties in ViTs without labeled data	Requires large-scale datasets and long training time
[6]	BYOL (Bootstrap Your Own Latent)	ImageNet	Linear evaluation accuracy	Introduces non-contrastive self-supervised learning using a teacher–student framework without negative samples	Lacks explicit modeling of local or patch-level representations
[7]	SimCLR	ImageNet	Contrastive loss / Top 1 accuracy	Learns strong global visual representations through instance-level contrastive learning	Requires large batch sizes and does not capture fine-grained spatial details
[8]	MoCo v3	ImageNet	Linear probing accuracy	Extends contrastive learning to Vision Transformers using momentum-based distillation	Focuses mainly on global representations without patch-level

					supervision
[9]	SwAV	ImageNet	Top 1 accuracy	Introduces clustering-based contrastive learning without pairwise instance comparison	Limited spatial awareness and lack of fine-grained feature learning
[10]	MAE (Masked Autoencoders)	ImageNet	Reconstruction loss / Linear evaluation	Learns strong patch-level representations through masked image modeling	Weak global semantic alignment across augmented views
[11]	BEiT	ImageNet	Linear probing accuracy	Applies BERT-style masked token prediction for patch-level image understanding	Does not explicitly enforce global invariance

3. PROPOSED MODEL ARCHITECTURE

A new method called a hybrid self-supervised learning framework is introduced to combine contrastive and non-contrastive learning to better learn detailed visual features from images. When an image is given, two different versions of it are created through data augmentation. Both versions go through the same Vision Transformer (ViT) model, which helps in extracting consistent features across different learning goals [8], [17].

The framework has two separate parts working at the same time. The first part of this system is called the contrastive branch. This part is inspired by MoCo v3 and SimCLR. It looks at the image. The contrastive branch uses something called the InfoNCE loss function. This helps create features that can tell things apart. These features stay the same even if the image is changed a bit [7], [24].

The second part is called the non-contrastive branch. The non-contrastive branch is based on BYOL and VICReg. This part uses a system. One part of the network is always. Updating itself. At the time, another part of the network keeps track of what it learned in the past. The non-contrastive branch and the contrastive branch are both parts of the system. This helps in learning without losing stability or running into problems with overfitting [19].

To get even more detailed features, the method also uses a type of self-supervision at the level of small parts of the image, called patches. These patches are made to follow goals that focus on both the meaning and the structure, inspire DINOv2 and iBOT. These features from patches add to the overall image features by capturing more detailed information about the local parts of the image. Finally, a special part of the framework combines the features from the whole image with those from the smaller patches [2], [12], as shown in **Figure 1**.

.

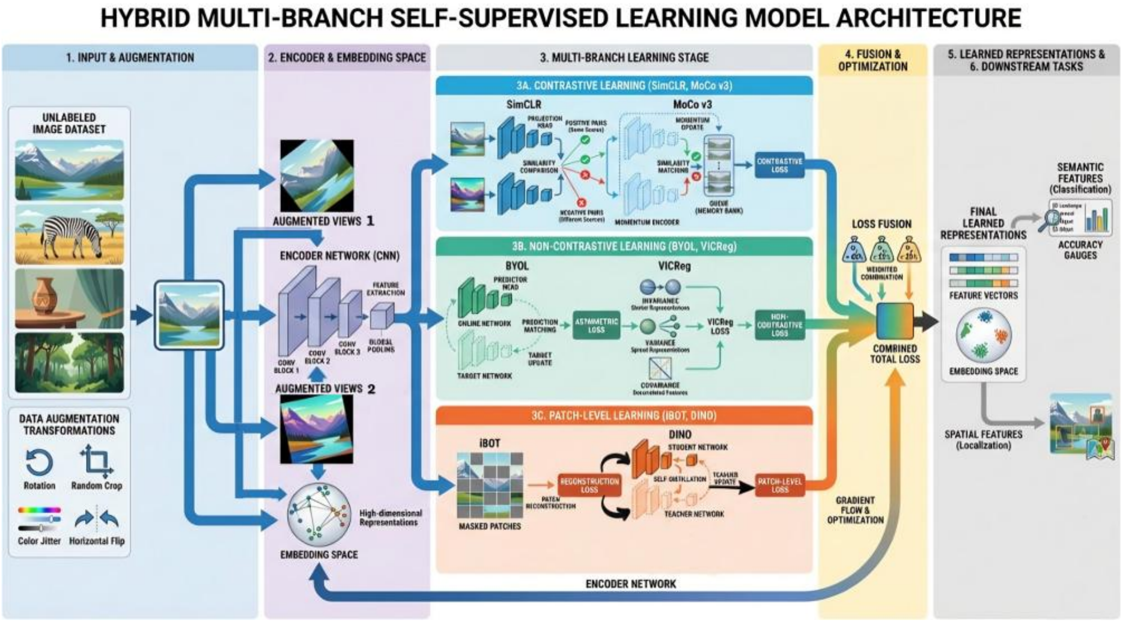
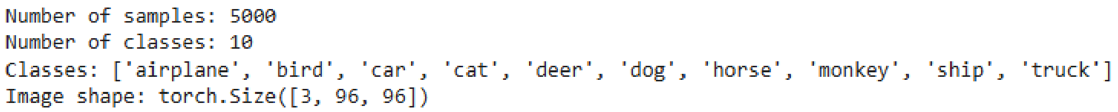


Figure 1. shows the model architecture

4. DATASET

Overview of dataset STL-10 [12] is an image recognition dataset inspired by the CIFAR-10 dataset with some improvements. With a corpus of 100,000 unlabeled images and 500 training images, this dataset is best for developing unsupervised feature learning, deep learning, and self-taught learning algorithms. Unlike CIFAR-10, the dataset has a higher resolution, which makes it a challenging benchmark for developing more scalable unsupervised learning methods.

There are three files: train_image.zips, test_images.zip and unlabeled_images.zip, 10 classes: airplane, bird, car, cat, deer, dog, horse, monkey, ship, truck. Images are 96x96 pixels, color, 500 training images (10 pre-defined folds), 800 test images per class, 100,000 unlabeled images for unsupervised learning. These examples are extracted from a similar but broader distribution of images. For instance, it contains other types of animals (bears, rabbits, etc.) and vehicles (trains, buses, etc.) in addition to the ones in the labeled set. Images were acquired from labeled examples on ImageNet, as shown in **Figure 2.**



```
Number of samples: 5000
Number of classes: 10
Classes: ['airplane', 'bird', 'car', 'cat', 'deer', 'dog', 'horse', 'monkey', 'ship', 'truck']
Image shape: torch.Size([3, 96, 96])
```

Figure 2. shows the number of samples and classes with their names

In data loading and preprocessing the data set was loaded by means of PyTorch's torchvision. datasets. STL10 class. For a preliminary study, raw images were converted to tensors without any data augmentation. A batch of raw images was displayed to inspect label distribution, image quality and class predominance, as provided in **Figure 3**.

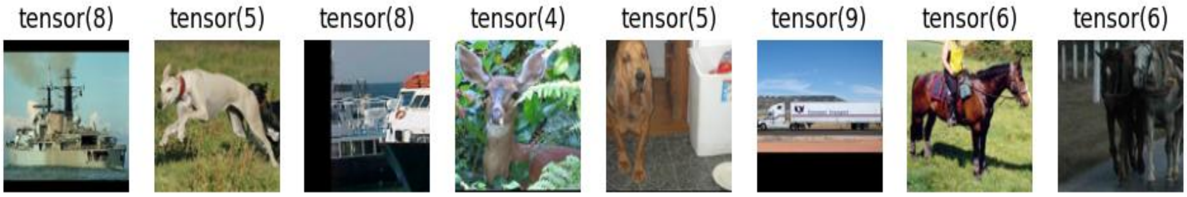
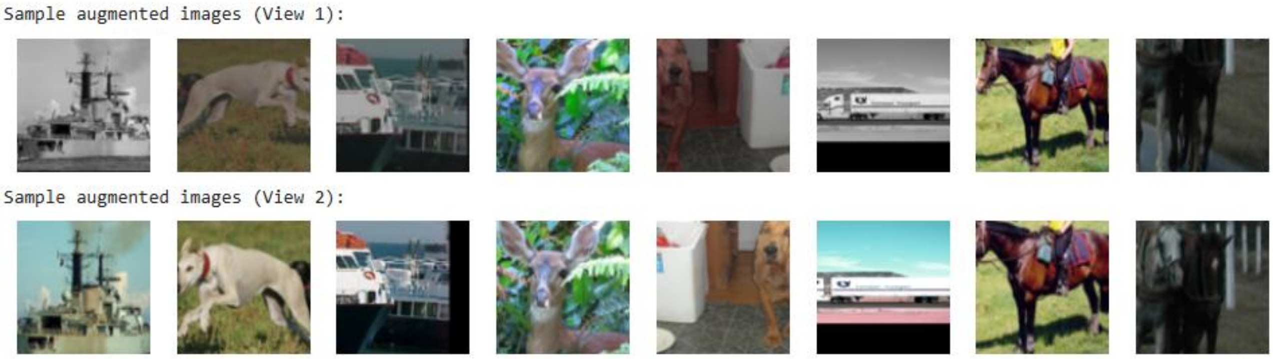


Figure 3. Sample Raw Images from STL-10 (Before Augmentation)

Use hybrid self-supervised augmentations to enable contrastive and non-contrastive self-supervised learning each image was transformed into two augmented views using the following augmentations: Random resized crop (scale: 0.5–1.0), Random horizontal flip (p=0.5), Color jitter (brightness, contrast, saturation, hue = 0.8∗s, s = 0.5), Random grayscale (p= 0.05) . These different augmentations simulate “different” “views” of the same image, which is crucial for BYOL, SimCLR, MoCo v3, and DINO components of the hybrid model as shown in **Figure 4**.



```
Sample augmented images (View 1):
```

```
Sample augmented images (View 2):
```

Figure 4. Sample SSL Augmented Images (View 1 and View 2)

Dataset visualization and pixel distribution to better understand the dataset and the effect of augmentations, pixel intensity histograms were plotted for raw and augmented images. These histograms demonstrate shifts in pixel distributions introduced by augmentations and ensure variability across views for effective self-supervised learning as shown in **Figure 5**.

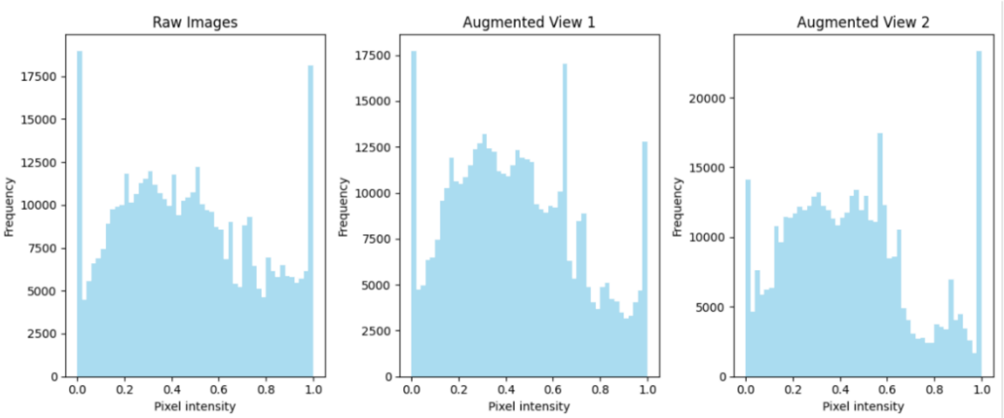


Figure 5. Pixel Intensity Histograms

In dataset statistics the dataset statistics are summarized as follows: Number of samples: 5,000 labeled images, 100,000 unlabeled images, Number of classes: 10, Image resolution: 96×96 pixels, 3 channels, Class distribution: Balanced across labeled data; unlabeled data contains additional variability (e.g., other animals and vehicles) as shown in **Figure 6**.

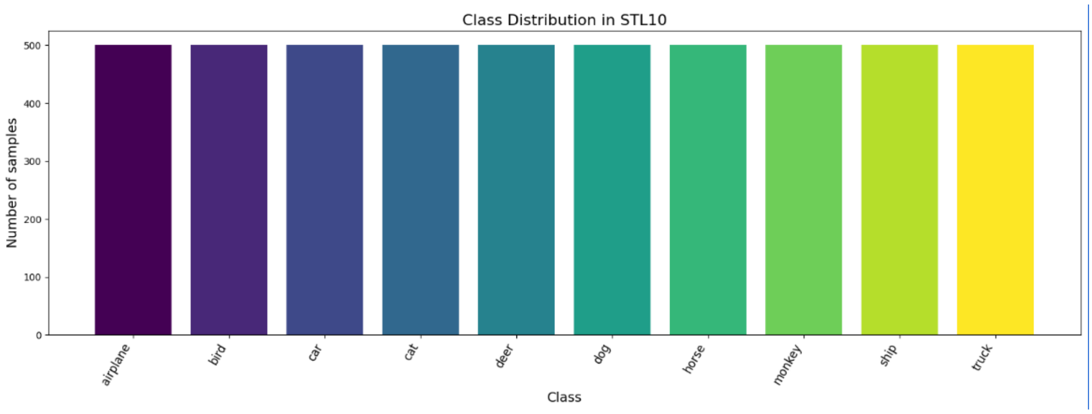


Figure 6. Class Distribution in STL-10

5. PROBLEM FORMULATION AND PROPOSED ARCHITECTURE

Self-supervised learning aims to learn meaningful visual representations from unlabeled data by exploiting intrinsic structures and augmentations [1], [11]. Let $X \subset \mathbb{R}^{H\times W\times C}$denote the space of natural images, where $H$and $W$are the height and width of the images, and $C$is the number of color channels. Given an unlabeled dataset $\mathcal{D} = \{x_i\}_{i=1}^{N}$, sampled from an unknown data distribution $p_{\text{data}}(x)$, The goal is to learn an encoder $f_\theta: X \to Z$parameterized by θ. Here, $Z \subset \mathbb{R}^d$is a low-dimensional representation space of dimension d, where $d \ll HWC$.

The learned representation $z = f_\theta(x) \in Z$is expected to satisfy two key properties. Semantic consistency in **Equation 1** requires that different stochastic augmentations $t_k, t_j \sim \mathcal{T}$of the same image $x$produce similar embeddings [16], [24]:

$$\mathbb{E}_{t_k,t_j\sim\mathcal{T}}[\text{sim}(f_\theta(t_k(x)), f_\theta(t_j(x)))] \approx 1 \qquad (1)$$

where $\text{sim}(\cdot,\cdot)$denotes the cosine similarity function, and $\mathcal{T}$is the set of all augmentation transformations applied during training. Discriminative power requires that the embeddings of distinct images x_i, x_j in **Equation 2** are dissimilar [9], [24]:

$$\mathbb{E}_{x_i,x_j\sim p_{\text{data}}, i\neq j}[\text{sim}(f_\theta(x_i), f_\theta(x_j))] \ll 1 \quad (2)$$

Despite the success of contrastive learning, it is limited by the InfoNCE objective embeddings as shown in **Equation 3 [**22**]**:

$$L_{\text{InfoNCE}} = -\mathbb{E}_{x\sim p_{\text{data}}} \log \frac{\exp\,(\text{sim}(f_\theta(t^+(x)), f_\theta(x))/\tau)}{\sum_{x^-\in\mathcal{N}} \exp\,(\text{sim}(f_\theta(t^-(x^-)), f_\theta(x))/\tau)} \qquad (3)$$

where $t^+(x)$denotes a positive augmented view of x, $t^-(x^-)$denotes negative samples from a set $\mathcal{N}$, and $\tau > 0$is a temperature parameter controlling the sharpness of the softmax. Strong invariance enforced by contrastive learning suppresses spatial information, i.e., $f_\theta(t(x)) \approx f_\theta(x)$for all $t \in \mathcal{T}$. Let $I_{\text{spatial}}: X \to S$denote the spatial information extractor; then the mutual information captured by contrastive embeddings, as shown in **Equation 4**:

$$I(f_\theta(X); I_{\text{spatial}}(X)) \ll I(X; I_{\text{spatial}}(X)) \qquad (4)$$

indicating a deficiency in preserving spatial features.

To formalize representation quality, we define a function using **Equations 5 and 6** [9], [24]:

$$Q(\theta) = \alpha Q_{\text{semantic}}(\theta) + \beta Q_{\text{spatial}}(\theta), \alpha, \beta > 0 \qquad (5)$$

Where,

$$Q_{\text{semantic}}(\theta) = \mathbb{E}_{(x,y)\sim p_{\text{task}}}[\mathcal{A}_{\text{cls}}(f_\theta(x), y)], Q_{\text{spatial}}(\theta) = \mathbb{E}_{(x,s)\sim p_{\text{spatial}}}[\mathcal{A}_{\text{loc}}(f_\theta(x), s)] \qquad (6)$$

Here, $\mathcal{A}_{\text{cls}}$and $\mathcal{A}_{\text{loc}}$denote classification and localization accuracy functions, respectively, $y$is a semantic label, and $s$represents spatial annotations (or pseudo-ground-truth spatial signals). Conventional contrastive methods satisfy **Equation 7** :

$$Q_{\text{spatial}}(\theta^*_{\text{contrastive}}) \ll \max_\theta\, Q_{\text{spatial}}(\theta) \qquad (7)$$

highlighting the suboptimal preservation of spatial information.

The research objective is therefore to find parameters in **Equation 8** :

$$\theta^* = \arg\max_\theta\, [Q_{\text{semantic}}(\theta) + Q_{\text{spatial}}(\theta)] \quad (8)$$

achieving both semantic and spatial representation quality without labeled data and within computational constraints.

To address these limitations, we propose a hybrid multi-branch architecture that unifies three self-supervised learning objectives. The total optimization problem is formulated, as shown in **Equation 9**.

$$\theta^* = \arg\min_{\theta} \quad L_{\text{total}}(\theta), \qquad L_{\text{total}}(\theta) = \sum_{i=1}^{3} \lambda_i \, L_i(\theta), \sum_{i=1}^{3} \lambda_i = 1 \qquad (9)$$

where $L_i(\theta)$denotes the loss associated with the i-th branch, and $\lambda_i$is its corresponding weight.

The **contrastive branch** combines the SimCLR model in **Equation 10** [16] and the MoCo v3 model in **Equations 11 and 12** [17], [25] objectives:

$$L_{\text{SimCLR}} = -\log \frac{\exp\,(\text{sim}(z_i,z_j)/\tau)}{\sum_{k\neq i} \exp\,(\text{sim}(z_i,z_k)/\tau)}, z_i = g(f_\theta(x_i)) \qquad (10)$$

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q, \quad L_{\text{MoCo}} = -\log \frac{\exp\,(q\cdot k^+/\tau)}{\exp\,(q\cdot k^+/\tau) + \sum_{k^-} \exp\,(q\cdot k^-/\tau)} \qquad (11)$$

$$L_{\text{contrastive}} = \alpha_1 L_{\text{SimCLR}} + \alpha_2 L_{\text{MoCo}}. \qquad (12)$$

Here, $f_\theta$is the encoder, $g$is the projection head, $z_i$is the projected embedding of the image x_i, $q$and $k^+, k^-$are query and positive/negative keys, $m$is the momentum coefficient, and $\tau$is the temperature. The combination weights $\alpha_1, \alpha_2$are typically 0.5.

The **non-contrastive branch** uses the BYOL model in **Equation 13** [18] and the VICReg model in **Equations 14 and 15** [19] objectives:

$$L_{\text{BYOL}} = \| \hat{h}_\theta(v) - \text{stopgrad}\left(\hat{g}_\xi(v')\right) \|_2^2 + \text{symmetrization} \qquad (13)$$

$$L_{\text{VICReg}} = \lambda s(Z) + \mu v(Z) + \nu c(Z) \qquad (14)$$

$$L_{\text{non-contrastive}} = \beta_1 L_{\text{BYOL}} + \beta_2 L_{\text{VICReg}} \qquad (15)$$

where $v, v'$are augmented views of the same image, $\hat{h}_\theta$and $\hat{g}_\xi$are predictor and target networks, $Z$is the batch of embeddings, $s(Z), v(Z), c(Z)$are invariance, variance, and covariance terms, and $\beta_1, \beta_2$are weighting coefficients.

The **patch-level branch** combines the iBOT model in **Equation 16** [21] and the DINO model in **Equations 17 and 18** [20] objectives:

$$L_{\text{iBOT}} = -\sum_{i\in M} p_t\,(i) \log p_s(i) \qquad (16) \qquad\qquad L_{\text{DINO}} = -\sum_{x} p_t\,(x) \log p_s(x)\,, p_t = \text{softmax}\left(\frac{g_t - c}{\tau_t}\right) \qquad (17)$$

$$L_{\text{patch}} = \gamma_1 L_{\text{iBOT}} + \gamma_2 L_{\text{DINO}} \qquad (18)$$

where $M$is the set of masked patches, $p_s, p_t$are student and teacher predictions, $g_t$is the teacher output, $c$is a centering vector, $\tau_t$is teacher temperature, and $\gamma_1, \gamma_2$are combination weights.

The **total loss** is therefore [23]:

$$L_{\text{total}} = \lambda_1 L_{\text{contrastive}} + \lambda_2 L_{\text{non-contrastive}} + \lambda_3 L_{\text{patch}} \qquad (19)$$

with typical values $\lambda_1 = 0.4, \lambda_2 = 0.3, \lambda_3 = 0.3$and branch combination weights $\alpha_i = \beta_i = \gamma_i = 0.5$. Optimization is performed using AdamW with a cosine learning rate schedule.

In conclusion, conventional contrastive methods are limited in fine-grained spatial reasoning. The proposed hybrid multi-branch framework in **Equation 19** balances semantic invariance and spatial sensitivity, enabling the learning of richer, more discriminative representations without labels [6], [10].

6. EXPERIMENTAL PROCESS

This section illustrates the experimental workflow of the proposed hybrid self-supervised learning (SSL)framework The overall approach follows a standard self-supervised visual representation learning pipeline and comprises six key steps: data preparation and augmentation, multi-view generation, backbone feature encoding, hybrid self-supervised representation learning, momentum-based target network updating, and representation evaluation [1], [11]. This study specifically applied the proposed unlabeled images from the STL-10 dataset that are first preprocessed and transformed using strong stochastic augmentations to generate multiple correlated views of each input sample. These views are then passed through a shared convolutional backbone and projection heads to extract latent representations. The proposed hybrid self-supervised model integrates contrastive and non-contrastive learning paradigms, inspired by SimCLR, MoCo v3, BYOL, and DINO, with the current implementation primarily leveraging a BYOL-based objective[18]. A momentum-updated target network is employed to stabilize training and prevent representation collapse. Finally, the learned representations are evaluated through downstream tasks to assess their effectiveness, generalization capability, and suitability for transfer learning in visual recognition applications[15] as shown in **Table 2**.

Table 2. Hybrid Self-Supervised of the STL-10 dataset.

Dataset Split	Purpose	Number Of Images	Image Size	Labels Use
Unlabeled set	Self-Supervised training	100,000	96 x 96 (RGB)	No
Training set	Downstream evaluation (Fine-tuning)	500(10 folds)	96 x 96 (RGB)	Yes
Test set	Performance evaluation	8000 (800 per class)	96 x 96 (RGB)	Yes

7. EXPERIMENTAL RESULT

The proposed hybrid self-supervised learning framework was trained on the unlabeled split of the STL-10 dataset using strong data augmentation and a momentum-based teacher–student architecture [10], [25]. The training process followed a fully self-supervised paradigm[1], [15], without utilizing any class labels. **Figure 7** illustrates a sample from data before the augmentation architecture of the hybrid SSL model, while **Figures 8, 9** presentsthe training workflow and feature learning process.

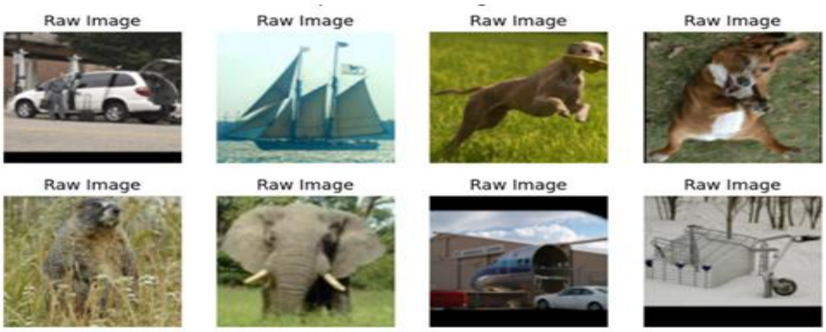


Figure 7. STL-10 Samples (Before Augmentation)

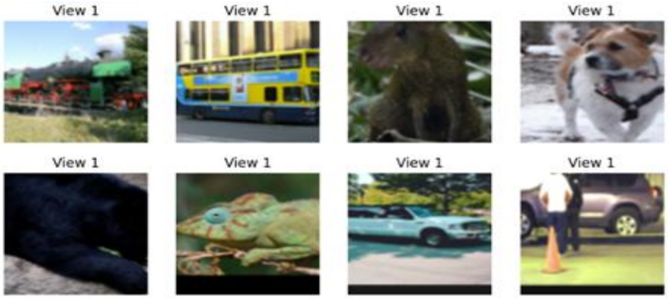


Figure 8. SSL Augmented Samples (View 1)

Figure 9. SSL Augmented Samples (View 2)

Figure 10 illustrates moderate classification performance with noticeable class-wise variations. The model achieves the highest accuracy in the monkey class (594 correct predictions) and shows relatively good performance on the ship (547) and car (392) classes. However, significant confusion exists between certain semantic categories, particularly between animals and between vehicles.

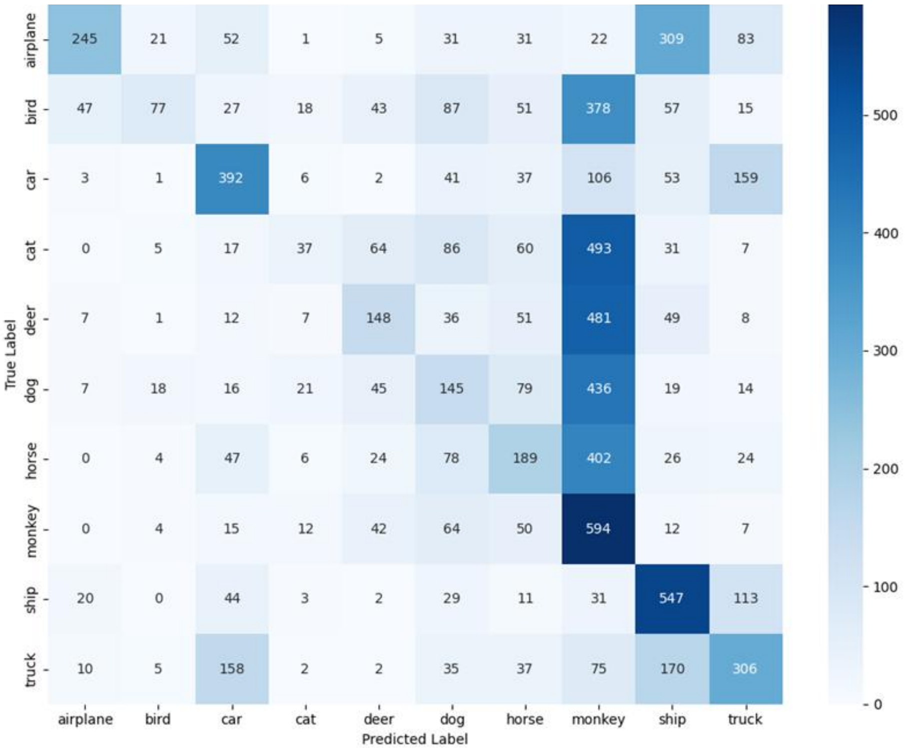


Figure 10. Confusion Matrix-Hybrid Self-Supervised Model

During training, the model converged stably over two epochs, achieving an average training loss of −0.8037 in Epoch 1 and −0.8837 in Epoch 2, as shown in the training log. The decreasing loss magnitude indicates effective alignment between online predictions and target network representations, confirming successful self-supervised optimization[24]. The negative loss values are expected due to the cosine similarity–based objective employed in the BYOL component of the hybrid framework[18].

8. CONCLUSION

This work introduces a unified hybrid self-supervised learning framework that jointly integrates contrastive and non-contrastive paradigms—namely SimCLR, MoCo v3, BYOL, and DINO—within a single architectural design to advance visual representation learning. By explicitly addressing key challenges such as limited labeled data availability and high intra-class visual variability, the proposed approach demonstrates that complementary self-supervised objectives can be effectively harmonized to improve both representation quality and training stability. A central contribution of this framework lies in the coordinated interaction between online and momentum-updated target networks, enabling balanced feature alignment while preserving representational diversity and variance. In addition, the use of carefully designed stochastic data augmentations facilitates the generation of diverse yet semantically consistent views, reinforcing robust feature learning. Experimental validation on the STL-10 dataset confirms stable convergence behavior and efficient acquisition of discriminative representations, highlighting the practicality of hybrid self-supervision as a reliable and computationally efficient alternative to single-paradigm self-supervised methods in large-scale unlabeled settings.

Despite these promising results, several research directions remain open, including the adaptive balancing of individual loss components and the scalability of the framework to larger backbone architectures and more diverse datasets. Future work will focus on extending the framework to multimodal scenarios, such as vision–language and vision–sensor learning, integrating dynamic objective-weighting strategies, and conducting more comprehensive downstream evaluations through linear probing and fine-tuning. These directions aim to further narrow the gap between invariant representation learning and spatially sensitive visual understanding.

REFERENCES

[1] T. Uelwer, J. Robine, S.S. Wagner, et al., A survey on self-supervised methods for visual representation learning, Mach. Learn. 114 (2025) 111.

[2] W. Qin, Y. Li, J. Zhang, X. Wen, J. Guo, Q. Guo, Attention-based hybrid contrastive learning for unsupervised person re-identification, Sci. Rep. 15(1) (2025) 13238.

[3] M. Kang, J. Kim, Enhancing self-supervised visual representation learning through adversarially generated examples, Neural Comput. & Applic. 37 (2025) 14613–14634.

[4] J. Nadine, Self-supervised learning principles challenges and emerging directions, Preprints (2025) 2025021894.

[5] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, P. Jiang, BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer, Proc. 28th ACM Int. Conf. Inf. Knowl. Manag. (2019) 1441–1450.

[6] S. Zhao, L. Zhu, X. Wang, et al., Slimmable networks for contrastive self-supervised learning, Int. J. Comput. Vis. 133 (2025) 1222–1237.

[7] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, arXiv preprint arXiv:2002.05709 (2020).

[8] X. Chen, S. Xie, K. He, An empirical study of training self-supervised vision transformers, arXiv preprint arXiv:2104.02057 (2021).

[9] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, G. Hinton, Big self-supervised models are strong semi-supervised learners, arXiv preprint arXiv:2006.10029 (2020).

[10] V.K. Reja, S. Goyal, K. Varghese, B. Ravindran, Q.P. Ha, Hybrid self-supervised learning-based architecture for construction progress monitoring, Autom. Constr. 153 (2023) 105225.

[11] A. Khan, S. AlBarri, M.A. Manzoor, Contrastive self-supervised learning: A survey on different architectures, Proc. 2nd Int. Conf. Artif. Intell. (2022) 1–6.

[12] G. Nagaraj, M.N. Rao, J.P. Wankhede, S.G. Rao, H.M. Abas, N. Gireesh, Self-supervised feature learning for robust image understanding in noisy and unstructured data, Proc. Int. Conf. Metaverse Curr. Trends Comput. (2025) 1-4.

[13] L. Shang, T. Wang, L. Gong, C. Wang, X. Zhou, Enhancing HLS performance prediction on FPGAs through multimodal representation learning, IEEE Embed. Syst. Lett. 16(4) (2024) 385-388.

[14] R.A. Jarvis, A perspective on range finding techniques for computer vision, IEEE Trans. Pattern Anal. Mach. Intell. 5(2) (1983) 122–139.

[15] I. Misra, L. van der Maaten, Self-supervised learning of pretext-invariant representations, Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (2020) 6707-6717.

[16] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, Proc. Int. Conf. Mach. Learn. 119 (2020) 1597-1607.

[17] X. Chen, S. Xie, K. He, An empirical study of training self-supervised vision transformers, Proc. IEEE/CVF Int. Conf. Comput. Vis. (2021) 9640-9649.

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

[19] A. Bardes, J. Ponce, Y. LeCun, VICReg: Variance-invariance-covariance regularization for self-supervised learning, Proc. Int. Conf. Learn. Represent. (2022).

[20] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, et al., Emerging properties in self-supervised vision transformers, Proc. IEEE/CVF Int. Conf. Comput. Vis. (2021) 9650-9660.

[21] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, et al., iBOT: Image BERT pre-training with online tokenizer, Proc. Int. Conf. Learn. Represent. (2022).

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

[23] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, Proc. Int. Conf. Learn. Represent. (2019).

[24] T. Wang, P. Isola, Understanding contrastive representation learning through alignment and uniformity on the hypersphere, Proc. Int. Conf. Mach. Learn. 119 (2020) 9929-9939.

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