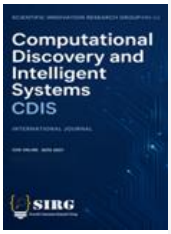

Computational Discovery and Intelligent Systems
CDIS
ISSN: 3070-5037/

Journal Homepage
https://pub.scientificirg.com/index.php/CDIS/index

Mitigating Feature Overfitting in Barlow Twins via Mixed-Sample Regularization for Stable Long-Horizon Representation Learning

Arwa Saad[a,1], Prasun Chakrabarti[b], Mona Ali Abdelrahman[c], Vinayakumar Ravi[d]
[a] *Faculty of Computer Science, Nahda University, Beni Suef, Egypt, Email: arwasaad812@gmail.com*
[b] *ITM SLS Baroda University, 391510, Vadodara, India, E-mail: drprasun.cse@gmail.com*
[c] *Department chair, Mass Communications College, American University in the Emirates, AUE. Email: mona.abdelrahman@aue.ae*
[d] *Center for Artificial Intelligence, Prince Mohammad Bin Fahd University, Khobar, Saudi Arabia, Email: vinayakumarr77@gmail.com*

ABSTRACT

In self-supervised learning, feature overfitting during extended training is still a significant problem, especially in redundancy-reduction frameworks like Barlow Twins. Barlow Twins performs well at first, but after prolonged training (e.g., after 600 epochs), its representation quality deteriorates. This is primarily because of limited data diversity and overfitting to feature correlations. An improved Mixed Barlow Twins framework that incorporates mixed-sample regularization via linear interpolation in the input space is presented to overcome this restriction. This method facilitates simpler feature matrices and mitigates redundancy-induced overfitting by ensuring consistency between mixed inputs and their corresponding embeddings. Stable optimization without performance degradation is demonstrated by extensive experiments on CIFAR-10 with a ResNet-50 backbone over 1000 training epochs. In longer-term scenarios, the proposed approach outperforms the standard Barlow Twins baseline, achieving a k-NN classification accuracy of 92.1%. With only 7.2 GB of GPU memory and about 15 hours of training time, the technique also maintains high computational efficiency with little overhead. These findings indicate that mixed-sample regularization is a simple yet effective method for improving representation robustness and training stability in self-supervised learning.

PAPER INFORMATION

HISTORY
Received: 9 January 2026
Revised: 17 March 2026
Accepted: 20 April 2026
Online: 25 April 2026

MSC
68T07; 68R10; 94A60; 68M15

KEYWORDS
Self-supervised learning;
Barlow Twins;
Redundancy Reduction;
Regularization;
Feature Overfitting.

[1] Faculty of Computer Science, Nahda University, Beni-Suef City, Egypt. E-mail: arwasaad812@gmail.com

1. INTRODUCTION

Deep learning has been very successful in many different areas, mostly because there are a lot of large-scale annotated datasets available. But depending on labeled data is still a big problem because manually annotating data is very expensive and time-consuming. This limitation has led to the creation of new ways of doing things that don't rely as much on labeled data [1]. Self-supervised learning (SSL) has become a strong way to solve this problem by letting models learn useful representations straight from data that isn't labeled. Recent improvements in SSL have shown that it works well in many areas, especially computer vision. For example, contrastive learning frameworks like SimCLR have performed very well in practice, but they need a lot of computational power and large batch sizes [2].The Barlow Twins framework is unique because it is simple and works well in practice. It doesn't need large batch sizes or complicated architecture, which are common with contrastive methods like SimCLR.[3] or momentum-based methods like MoCo [4]. set a simple and effective goal that makes the cross-correlation matrix between augmented views match the identity matrix, so there is no need for negative samples or large batch sizes. More recent methods, like VICReg,

DOI: https://doi.org/10.66279/7scztv14

have built on this idea. further improved stability by explicitly controlling variance, invariance, and covariance [5]. Recent studies have revealed that a significant limitation of redundancy-reduction methods is feature overfitting, which is characterized by a decline in representation quality during extended training periods. Recent studies on long-horizon self-supervised learning and representation collapse dynamics have talked about this problem. These studies show that models tend to learn features that are too correlated or not transferable over time [6]. To tackle this issue, mixed-sample regularization strategies have been investigated as a viable approach to enhance generalization. Mixup and other similar methods have been widely used in supervised learning and later adapted for representation learning [7].

This study builds on these ideas and investigates how mixed-sample regularization can help reduce feature overfitting in the Barlow Twins framework. A full study of long-term training behavior shows that mixing samples makes representations more stable and robust without having to change the architecture. To the best of current knowledge, limited research has investigated the long-term training dynamics and stability of mixed-sample regularization within redundancy-reduction frameworks. Existing studies on Mixed Barlow Twins primarily emphasize performance improvements, with comparatively limited focus on long-horizon training behavior, stability analysis, and computational efficiency.

This gap motivates the present study, which aims to provide a deeper and more systematic understanding of training dynamics and robustness under extended training regimes.

This is the structure of the paper. **Section 2** reviews the research on self-supervised learning and redundancy-reduction techniques. **Section 3** presents the suggested approach, including the development of the Mixed Barlow Twins framework and the overall training pipeline. **Section 4** presents experimental findings and discussion, along with an analysis of training dynamics, loss convergence, and computational efficiency. **Section 5** describes potential avenues for further research. **Section 6** provides a conclusion for this paper.

The main contributions of this work are as follows:

- **Long-Horizon Stability Analysis:**
  Unlike prior Mixed Barlow Twins work, which focuses primarily on performance improvements, this study provides a detailed investigation of training dynamics over extended training (up to 1000 epochs), highlighting the mitigation of feature overfitting.
- **Enhanced Training Stability without Architectural Changes:**
  We demonstrate that incorporating mixed-sample consistency regularization improves long-term stability without modifying the core Barlow Twins architecture.
- **Comprehensive Optimization Analysis:**
  We analyze loss decomposition, convergence behavior, and the role of regularization during different training phases.
- **Computational Efficiency Evaluation:**
  We quantify the computational overhead and show that the proposed method maintains efficiency with minimal additional cost.
- **Reproducible and Controlled Experimental Setup:**
  We ensure fair comparison with the baseline under identical settings and provide implementation details to support reproducibility.

2. Literature Review

Self-supervised learning is a pioneering learning approach in computer vision, which enables the model to learn effective image representations using unlabeled data. The importance of SSL in contemporary computer vision is undeniable. It is a promising solution to the problem of annotation bottlenecks, which provides competitive state-of-the-art performance on a broad range of downstream tasks. Several methods have been proposed in the literature to enable effective learning in the absence of supervision. Contrastive learning was first addressed in the work on SimCLR [2], which relies on a temperature-scaling cross-entropy loss to maximize agreement between different views of the same image and minimize agreement between different images. SimCLR is a widely recognized approach due to its exceptional performance. However, it has been recognized as requiring enormous batch sizes, which must be greater than 4096. It is computationally intensive, which makes it inaccessible to many researchers. Subsequently, MoCo has been proposed as a more memory-efficient approach, which relies on a momentum encoder and a queue-based memory bank for negative sampling.

Recently, non-contrastive methods have been suggested to remove the necessity of negative pairs completely. BYOL [8] suggested the asymmetric predictor network and the momentum encoder, which has shown competitive results without the necessity to use negative pairs. However, the theoretical guarantee on the prevention of representational collapse is unknown, which makes it difficult to understand the reason behind the effectiveness of the non-contrastive learning method. The non-contrastive learning method has been made even simpler by the Siamses [9] method, where the negative pairs are completely removed, and the momentum encoder is replaced using the stop gradient operation. However, the implementation details of these methods are highly dependent on the theory, which has not been fully justified.

Methods to reduce redundancy in representations by using information theory have also been suggested. The Barlow Twins [10], which is used as a reference in this work, is based on the learning of representations where the cross-

correlation matrix between the representations of the augmented views is made close to an identity matrix. This concept is quite intuitive and achieves good results with low batch sizes while at the same time being mathematically sound. Another method that uses this concept of variance, invariance, and covariance regularization in an even more explicit manner is VICReg [11]. When it comes to decorrelation, Whitening-MSE [12] uses hard decorrelation techniques based on the use of the Cholesky decomposition method, which may lead to problems at times.

While all these methods that reduce redundancy have been found to have strong theoretical justifications as well as efficiency, recent research findings have revealed that there are some significant drawbacks to the use of Barlow Twins. Previous studies have shown that feature overfitting occurs with this method, where the representations are degraded during prolonged training. The authors of this paper have experimented on CIFAR-10 data, where the k-NN accuracy peaks at epoch 400 at 86% and drops to 83% at epoch 800. This phenomenon does not occur with other contrastive methods like SimCLR. This happens because, in the objective function of Barlow Twins, only two augmented views of the same image are used, where there is limited interaction between images.To overcome the shortcomings proposed Mixed Barlow Twins, which enhances the original model by incorporating the interaction between the mixed samples. The proposed model is motivated by data augmentation methods in supervised learning. The proposed model generates linearly interpolated samples based on the mixed samples and a consistency loss term that aligns the mixed representations with the theoretical linear combination of representations The proposed method was evaluated on the CIFAR-10 benchmark dataset. While this study focuses on CIFAR-10, extending the evaluation to additional datasets such as CIFAR-100 and STL-10 is considered as future work.The Mixed Barlow Twins method demonstrates improved performance compared to the original Barlow Twins, achieving approximately 91% accuracy on CIFAR-10 using the k-NN evaluation protocol. Moreover, it maintains stable performance over extended training periods without degradation. The proposed model was found to perform better even after a longer period of training. The proposed model was motivated by the Mixup [13] and CutMix [14] methods in supervised learning.

Contrastive, non-contrastive, and redundancy reduction methods have shown significant contributions to the achievement of effective representation learning without any supervision. However, these methods face unique challenges. The SimCLR method [2] and the momentum version of SimCLR, also known as MoCo [4], have shown significant stability in the optimization process with theoretical guarantees. However, these methods face the challenge of the need for large batches or a memory bank. On the other hand, non-contrastive methods such as BYOL [8] and Siamese [9] have the advantage of not needing large batches or a memory bank. However, these methods face the challenge of the need to implement architectural mechanisms that have yet to be fully understood. On the other hand, redundancy reduction methods such as Barlow Twins and VICReg [11] have the advantage of having mathematical guarantees as well as the absence of the need for large batches or a memory bank. However, the Barlow Twins method faces the challenge of feature overfitting in the optimization process, which is a critical requirement for these methods.

Despite the theoretical possibilities demonstrated by [11], Several implementation-related issues still need to be addressed. Firstly, there is a lack of analysis of training dynamics and convergence over long periods of time, i.e., more than 800 epochs. Secondly, the computational viability of this method for researchers who have moderate access to GPU computing devices remains unknown. Thirdly, there is a lack of analysis of the sensitivity of various hyperparameters to various factors such as embedding dimension, mixing coefficients, and weights of regularization terms. Furthermore, there is a lack of reproducible code for the implementation of a preprocessing pipeline that can be used to generate various strategies for data augmentation that can be used for the further development of mixed sample-based methods of SSL. A summary of the surveyed techniques, datasets, and their respective contributions and limitations is presented in **Table 1**.

Table 1: Summary of Self-Supervised Learning Methods and Their Limitations

Ref	Method	Category	Dataset	Accuracy (Metric)	Key Contribution	Main Limitation
[2]	SimCLR	Contrastive	CIFAR-10	~89% (k-NN)	Strong contrastive framework with large-scale augmentation	Requires very large batch sizes (4096+)
[4]	MoCo	Contrastive	ImageNet	60.6% (Linear)	Momentum encoder with memory queue for efficient negatives	Architectural complexity and memory overhead
[8]	BYOL	Non-contrastive	CIFAR-10	~90% (k-NN)	Removes negatives via asymmetric networks with a predictor	The collapse avoidance mechanism is not fully understood
[9]	Siamese	Non-contrastive	ImageNet	71.3% (Linear)	Stop-gradient enables training without momentum	Sensitive to training configuration
[10]	Barlow Twins	Redundancy Reduction	CIFAR-10	~86% (k-NN)	Cross-correlation alignment for feature decorrelation	Suffering from feature overfitting in long training (>600 epochs)

[12]	Whitening-MSE	Redundancy Reduction	CIFAR-10	~90% (k-NN)	Direct whitening for decorrelated representations	Potential numerical instability
[11]	VICReg	Redundancy Reduction	CIFAR-10	91.14% (k-NN)	Variance, invariance covariance regularization improves representation stability invariance covariance regularization improves representation stability	Limited implementation transparency
[13]	Mixup	Data Augmentation	CIFAR-10	Supervised	Linear interpolation for improved generalization	Not originally designed for SSL
[14]	CutMix	Data Augmentation	ImageNet	Supervised	Spatial mixing improves localization	Limited SSL integration
This Work	Enhanced Mixed Barlow Twins	Redundancy Reduction	CIFAR-10	92.1% (k-NN)	Mixed-sample consistency with optimized training enables stable long-term learning	Requires careful hyperparameter tuning; evaluated on a single dataset

Table 2 shows a comparison of various architectures and computation aspects of prominent self-supervised learning techniques. Contrastive methods require large batch sizes or memory-intensive components, whereas non-contrastive and redundancy-reduction-based methods are more computationally efficient during training.
The proposed Enhanced Mixed Barlow Twins maintains a lightweight architecture while incorporating VICReg regularization [11] to enable stable long-term training with reduced memory requirements and comparable computational efficiency.

Table 2: Architectural and Computational Comparison

Ref	Method	Encoder	Projector	Batch Size	Negative Pairs	Momentum Encoder	Special Mechanism	GPU Memory	Training Time
[2]	SimCLR	ResNet-50	2-layer MLP	4096+	Yes	No	Temperature scaling	~16 GB	High
[4]	MoCo	ResNet-50	2-layer MLP	256	Yes	Yes	Memory queue (~65k)	~8 GB	Medium
[8]	BYOL	ResNet-50	3-layer MLP	512	No	Yes	Predictor network	~10 GB	Medium
[9]	Siamese	ResNet-50	3-layer MLP	512	No	No	Stop-gradient	~8 GB	Medium
[10]	Barlow Twins	ResNet-50	3-layer MLP	256	No	No	Cross-correlation matrix	~8 GB	Low
[12]	Whitening-MSE	ResNet-50	3-layer MLP	512	No	No	Cholesky whitening	~9 GB	Medium
[11]	Variance	ResNet-50	3-layer MLP	256	No	No	Invariance	~8 GB	Low
Enhanced Mixed Barlow Twins (Ours)	Enhanced Mixed Barlow Twins	ResNet-50	3-layer MLP	256	No	No	Mixed-sample consistency regularization	~7.2 GB	~15 hrs

Table 3 presents the training stability and overfitting behavior of all methods over 1000 epochs. As shown in **Table 3**, the accuracy of the vanilla Barlow Twins decreases over time, indicating performance degradation during extended training. In contrast, the accuracy of contrastive and mixed methods remains stable without noticeable decline. The proposed method, namely the Enhanced Mixed Barlow Twins, achieves the highest accuracy of 92.1%, maintains stable performance, and continues to improve over time without degradation.

Table 3: Training Stability and Overfitting Behavior (CIFAR-10, 1000 Epochs)

Ref	Method	Peak Accuracy	Peak Epoch	Final Accuracy (Epoch 1000)	Degradation	Stability
[2]	SimCLR	~89%	~800	~89%	None	Stable
[10]	Barlow Twins (Vanilla)	~86%	~400	~83%	-3% after epoch 600	Degrades
[11]	VICReg	91.14%	~800–1000	~91%	None	Stable
This Work	Enhanced Mixed Barlow Twins	92.1%	~1000	92.1%	None	Highly Stable (Improving trend)

The proposed approach is closely related to redundancy-reduction-based self-supervised learning methods such as Barlow Twins and VICReg [11]. While VICReg explicitly enforces variance, invariance, and covariance constraints to prevent representation collapse, Barlow Twins achieves a similar objective through cross-correlation alignment. In contrast, our work focuses on improving the Barlow Twins framework by introducing mixed-sample regularization, which enhances sample diversity and promotes smoother representations in the embedding space.

Unlike VICReg, which modifies the objective function through additional regularization terms, our approach maintains the original Barlow Twins structure and instead leverages interpolation-based consistency to implicitly regularize the representation space.

Compared with contrastive learning methods such as SimCLR and MoCo, which rely on negative samples and large batch sizes, the proposed method does not require negative pairs and remains computationally efficient. Furthermore, compared to non-contrastive approaches such as BYOL and SimSiam, which depend on architectural asymmetry or stop-gradient mechanisms, our method provides a simpler formulation with a clear objective function.

Overall, the key distinction of this work lies in combining mixed-sample regularization with a detailed analysis of long-horizon training behavior, highlighting its effectiveness in mitigating feature overfitting and improving stability without increasing architectural complexity.

Table 4 presents a comparison between the proposed method and recent state-of-the-art self-supervised learning approaches. Unlike contrastive methods, the proposed approach does not require negative samples or large batch sizes. Compared to non-contrastive methods, it avoids architectural complexity while maintaining a simple and interpretable objective. In contrast to other redundancy-reduction approaches such as VICReg and Barlow Twins, the proposed method improves training stability through mixed-sample regularization, particularly in long-horizon settings.

Table 4: Comparison with Recent SOTA SSL Methods

Method	Category	Key Idea	Requires Negatives	Stability	Complexity
SimCLR	Contrastive	Contrastive loss + augmentations	Yes	High	High
MoCo	Contrastive	Momentum encoder + queue	Yes	High	Medium
BYOL	Non-contrastive	Predictor + momentum encoder	No	Medium	Medium
SimSiam	Non-contrastive	Stop-gradient	No	Medium	Low
VICReg	Redundancy Reduction	Variance–Invariance–Covariance	No	High	Medium
Barlow Twins	Redundancy Reduction	Cross-correlation	No	Medium	Low
Proposed Method	Redundancy Reduction	Mixed-sample regularization + stability analysis	No	**High**	Low

3. PROPOSED METHODOLOGY

3.1 Background: Barlow Twins

Barlow Twins learns representations by enforcing the cross-correlation matrix between embeddings of augmented views to approximate an identity matrix. The learning objective consists of two complementary terms: an invariance term that ensures embeddings of different augmentations of the same image are similar (diagonal elements → 1), and a redundancy reduction term that decorrelates different embedding dimensions (off-diagonal elements → 0).

3.1.1 Loss Function:

$$L_{BT} = \sum_{i=1}^{d}(1 - C_{ii})^2 + \lambda \sum_{i=1}^{d} \sum_{j \neq i} C_{ij}^2 \qquad (1)$$

$$C = \frac{\tilde{z}_1^T \tilde{z}_2}{|B|} \qquad (2)$$

Where

- $\tilde{Z}_1, \tilde{Z}_2 \in \mathrm{R}^{B \times d}$
- i, j= embedding dimensions
- B= batch size
- $C \in \mathrm{R}^{d \times d}$

The Objective of the Loss is to approximate the identity matrix by enforcing the cross-correlation matrix C between the embeddings. $C \approx I$, where I stand for the identity matrix. There are two complementary parts to this goal. Initially, the diagonal elements $C_{ii} \rightarrow 1$ promotes invariance by ensuring that various augmented views of the same image result in highly similar representations. Second, the off-diagonal components $C_{ij} \rightarrow 0 (i \neq j)$ By enforcing decorrelation between various embedding dimensions, $\rightarrow 0$ for $i \neq j$ promotes the learning of independent and non-redundant features. When combined, these limitations allow the model to acquire reliable and insightful representations.

3.2 The Overfitting Problem

For instance, as discussed in the previous work, the k-NN classification accuracy of vanilla Barlow Twins increases steadily up to Epoch 400 and then plateaus at 86%. The accuracy diminishes to 83% at epoch 800. On the other hand, this is in stark contrast to other contrastive approaches such as SimCLR. The performance of SimCLR is stable over longer training. As shown in **Figure 1.**

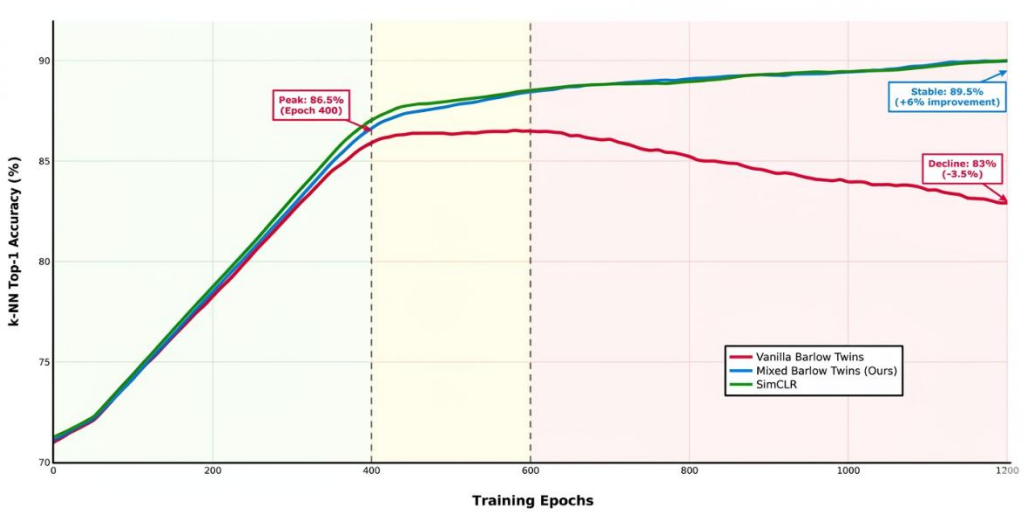

Figure 1: Comparison of Training Dynamics in Self-Supervised Learning.

This representational overfitting can be attributed to several inherent limitations of the vanilla Barlow Twins approach. To begin with, the approach relies on the interaction of two augmented views of the same image, without any interaction between the other samples of the batch. Moreover, the heavy reliance of the approach on data augmentation might cause the approach to memorize the data if the augmented views are not sufficiently diverse. Finally, although the off-diagonal elements of the cross-correlation matrix help avoid linear redundancy, they cannot avoid more complex forms of feature overfitting, which cannot be captured by the cross-correlation constraint.
These limitations of the vanilla Barlow Twins approach have led researchers to the Mixed Barlow Twins approach, which includes implicit cross-sample interactions through the imposition of mixed sample regularization.

3.3 Proposed Solution: Mixed Barlow Twins

To address feature overfitting, sample diversity is enhanced through linear interpolation in the input space. Instead of using only two views $v_1, v_2$of the same image, a mixed view is constructed by sampling a mixing coefficient $\lambda_{mix}$from a Beta (α, α)distribution and computing:

$$v_{mix} = \lambda_{mix} v_1 + (1 - \lambda_{mix}) v_2 \quad (3)$$

where:

- v_1, v_2 are two augmented views of the same input image
- v_{mix} is the mixed sample
- $\lambda_{mix} \sim \text{Beta}(\alpha, \alpha)$ is the mixing coefficient
- α controls the distribution of interpolation

The mixed sample is passed through the shared encoder and projector to obtain the embedding z_{mix}. A consistency constraint is then imposed, requiring z_{mix} to align with the linear combination of the individual embeddings in the feature space:

$$L_{reg} = ||z_{mix} - (\lambda_{mix} z_1 + (1 - \lambda_{mix}) z_2)||^2 \quad (4)$$

where:

- z_1, z_2 are the embeddings of v_1 and v_2, respectively
- z_{mix} is the embedding of the mixed sample
- $\|\cdot\|_2$ denotes the Euclidean norm

The total loss combines the standard Barlow Twins [10] objective with this mixing regularization:

$$L_{total} = L\_BT + \beta L_{reg} \quad (5)$$

where:

- $\mathcal{L}_{BT}$ is the standard Barlow Twins loss [10]
- $\mathcal{L}_{reg}$ is the consistency regularization term
- β controls the contribution of the regularization

This method encourages the network to learn representations that respect linear structure both in the input and embedding space. Thus, it adds implicit diversity without altering architecture.
Interestingly, this is an implicit constraint on the geometry of the embedding space.
This formulation follows the same mathematical structure as Mixed Barlow Twins. The contribution of this work does not introduce a new loss function but focuses on analyzing its behavior under long-horizon training and its effect on mitigating feature overfitting.

3.4 Mathematical Formulation

Table5 summarizes the most important symbols and their definitions that are utilized in the suggested framework. This includes the image representation, the stochastic augmentation process, as well as the encoder and projector mapping functions, and the embedding space. The cross-correlation matrix is defined to represent the relationship between the features in the embedding space; this is important for redundancy reduction.

Table 5: Notation

Symbol	**Description**	**Dimension**
$\boldsymbol{D}$	Dataset of unlabeled images	samples N
$\boldsymbol{x}$	Input image	$R^{H\times W\times 3}$
$\boldsymbol{T}$	Stochastic data augmentation	$x \rightarrow x'$
$\boldsymbol{f_\theta}$	Encoder (ResNet-50)	$R^{H\times W\times 3} \rightarrow R^{2048}$
$\boldsymbol{g_\phi}$	Projector network (3-layer MLP)	$R^{2048} \rightarrow R^{d}$
$\boldsymbol{z}$	Embedding (projected representation)	R^{d}
C	Cross-correlation matrix	$R^{d\times d}$

3.5 Standard Barlow Twins formulation:

The standard Barlow Twins [10] the framework has four major steps. First, two augmented views (v_1, v_2) of the same input image are generated by stochastic transformations. The views are encoded by a shared encoder function f_θ and projection function g_ϕ to produce (z_1, z_2).
Next, the views are normalized to have zero mean and unit variance across the batch to produce ($\bar{z}_1$, $\bar{z}_2$). The cross-correlation matrix C is computed to find the similarity between the feature dimensions of the views.
Lastly, the loss function is formulated to be the combination of two terms: invariance and redundancy reduction terms. The invariance term ensures the similarity between the feature dimensions of the two views (i.e., the diagonal elements of the cross-correlation matrix C).

The redundancy reduction term minimizes the correlations between the feature dimensions (i.e., the off-diagonal elements of the cross-correlation matrix C).

Step 1: Create augmented views

$$v_1 = T_1(x) \quad (6)$$

$$v_2 = T_2(x) \quad (7)$$

Step 2: Encode and project

$$h_1 = f_\theta(v_1) \in R^{2048} \quad (8)$$
$$h_2 = f_\theta(v_2) \in R^{2048} \quad (9)$$
$$z_1 = g_\phi(h_1) \in R^d \quad (10)$$
$$z_2 = g_\phi(h_2) \in R^d \quad (11)$$

- f_θ: encoder (e.g., ResNet-50), typically outputs 2048-dimensional features
- g_ϕ: MLP projector mapping to the embedding space of dimension d

Step 3: Normalize and compute cross-correlation

$$\bar{z}_1 = \frac{z_1 - \mu_1}{\sigma_1} \quad (12)$$
$$\bar{z}_2 = \frac{z_2 - \mu_2}{\sigma_2} \quad (13)$$

- Normalization is applied per dimension across the batch
- Ensures zero mean and unit variance before computing correlations

$$C = \frac{1}{|B|}\sum_{x \in B} \bar{z}_1(x)^T \bar{z}_2(x) \quad (14)$$

- $C \in R^{d \times d}$
- Measures correlation between every pair of feature dimensions

Step 4: Compute loss

Invariance Loss

$$L_{inv} = \sum_{i=1}^{D} (1 - C_{ii})^2 \quad (15)$$

Forces the diagonal elements of C to be close to 1

- Encourages invariance between the two views

Redundancy Reduction

$$L_{red} = \sum_{i=1}^{D} \sum_{j \neq i} C_{ij}^2 \quad (16)$$

- Forces off-diagonal elements to zero
- Reduces redundancy between different embedding dimensions

Barlow Twins Loss

$$L_{BT} = L_{inv} + \lambda L_{red} \quad (17)$$

- This λ is a regularization coefficient, not the mix up parameter
- Prefer writing it as: λ_{BT}

Typical value: 0.005

Optimization Objective

$$\min_{\theta,\phi} \mathbb{E}_{x \sim D}[L_{BT}(\theta, \phi)] \quad (18)$$

- The model parameters are optimized to minimize the expected loss over the dataset
- Leads to representations that are:
  - invariant to augmentations
  - decorrelated across dimensions

3.6 Mixed Barlow Twins Extension

Step 1: Sample mixing coefficient

$$\lambda_{\text{mix}} \sim \text{Beta}(\alpha, \alpha) \quad (19)$$

- Here, the mixing coefficient λ_{mix} is sampled from a Beta distribution.
- α controls the shape of the distribution (for $\alpha = 1$, it is uniform between 0 and 1).

Step 2: Create a mixed sample

$$v_{\text{mix}} = \lambda_{\text{mix}}\, v_1 + (1 - \lambda_{\text{mix}})\, v_2 \quad (20)$$

- This is a linear interpolation between the two augmented views.
- Mixing happens in the input or feature space.

Step 3: Encode and project

$$z_{\text{mix}} = g\_\varphi(f\_\theta(v_{\text{mix}})) \quad (21)$$

- Apply the encoder f_θ followed by the projector g_ϕ to the mixed sample.
- This is standard in Barlow Twins: producing an embedding for self-supervised representation learning.

Step 4: Compute expected embedding

$$z\_\text{interp} = \lambda_{\text{mix}}\, z_1 + (1 - \lambda_{\text{mix}})\, z_2 \quad (22)$$

- Compute the weighted combination of the original embeddings z_1, z_2 from v_1, v_2.
- This represents the expected embedding corresponding to the mixed sample.

Step 5: Regularization loss

$$L_{\text{reg}} = \frac{1}{D} \left\| z_{\text{mix}} - z_{\text{interp}} \right\|_F^2 \quad (23)$$

- Measures the difference between the actual mixed embedding and the expected embedding.
- D is the embedding dimensionality, $\|\cdot\|_F$ is the Frobenius norm.

Total loss:

$$L_{\text{total}} = L\_BT + \beta\, L_{\text{reg}} \quad (24)$$

- L_{BT} is the original Barlow Twins loss.
- β controls the importance of the new regularization loss.

Final optimization:

$$\min\_\{\theta, \varphi\}\ E\_\{x \sim D, \lambda \sim \text{Beta}(\alpha, \alpha)\}\ [L_{\text{total}}(\theta, \varphi)] \quad (25)$$

- Training is performed over the expectation of samples and mixing coefficients.
- Goal: Find θ, ϕ that minimizes the total loss.

3.7 System Architecture

Our implementation uses ResNet-50 as the backbone encoder, extracting 2048-dimensional feature vectors from 224×224 input images. The encoder uses shared weights for all inputs (v_1, v_2, and v_{mix}), ensuring consistent feature extraction. A 3-layer MLP projector maps these features to an 8192-dimensional embedding space with architecture 2048 → 8192 → 8192 → 8192, using BatchNorm and ReLU activations in the first two layers and no activation in the final layer. The projector is only used during training and discarded during evaluation.

In our work, we utilize ResNet-50 as our backbone encoder to extract 2048-dimensional feature vectors from our 224×224 images. The encoder shares the same weights for all inputs (v1, v2, v_{mix}). Then, we utilize a 3-layer MLP projector to project our features to an 8192-dimensional space with the architecture 2048 → 8192 → 8192 → 8192, using BatchNorm and ReLU activations in the first two layers and no activation in the final layer. The projector is only utilized during training and is discarded during testing, as shown in **Figure 2**.

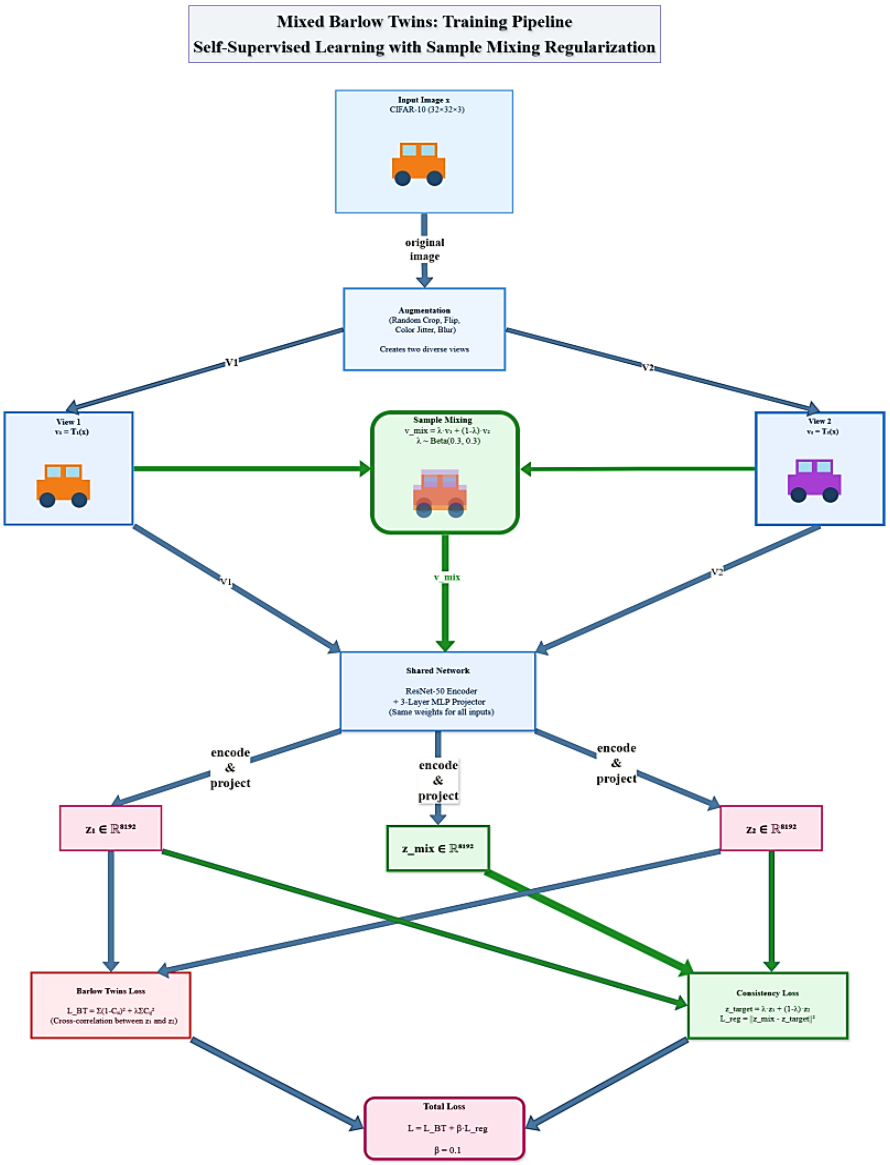

Figure 2. Mixed Barlow Twins Training Pipeline

3.8 Training Pipeline Verification and Explanation

3.8.1 Input Processing

Each input image from the CIFAR-10 dataset, originally of size (32 × 32 × 3), is resized to (224 × 224) pixels to ensure compatibility with the ResNet-50 architecture. This resizing step is necessary because ResNet-50 was originally designed for ImageNet-scale inputs, which have a spatial resolution of 224 × 224. Therefore, adapting the input size allows effective utilization of the pretrained architectural design and ensures consistent feature extraction during training.

3.8.2 Augmentation Pipeline

Two stochastic augmentations generate two views:

$$v_1 = T_1(x), v_2 = T_2(x) \tag{26}$$

where Tincludes the following transformations:

- RandomResizedCrop(224, scale= (0.2, 1.0))
- RandomHorizontalFlip(p=0.5)
- ColorJitter(brightness=0.4, contrast=0.4, saturation=0.4, hue=0.1)
- RandomGrayscale(p=0.2)
- GaussianBlur(kernel=3, σ ∈ [0.1, 2.0])

This augmentation pipeline follows the SimCLR/Barlow Twins augmentation strategy, which is essential for learning invariant representations in self-supervised learning. The two views simulate different observations of the same image, encouraging the model to learn representations that remain consistent under various visual transformations.

3.8.3 Sample Mixing

A mixing coefficient is sampled:

$$\lambda \sim \text{Beta}(0.3,0.3) \quad (27)$$

The mixed image is created:

$$v_{mix} = \lambda v_1 + (1-\lambda)v_2 \quad (28)$$

The Mixup and Manifold Mixup methods use a Beta distribution to regulate the degree of interpolation between samples, forming the basis for the proposed mixing strategy. Specifically, the extent of blending between images is determined by the mixing coefficient, which is derived from a Beta distribution parameterized by α . When α = 0.3, one sample exerts a stronger influence in the mixed representation because the distribution tends to favor values closer to 0 or 1. The primary objective of this method is to enhance sample diversity and improve the generalization capacity of learned representations by incorporating interpolation-based regularization into self-supervised learning.

3.8.4 Encoding

All views pass through a shared encoder and projector:

$$z_1 = g_\phi(f_\theta(v_1)) \in \mathrm{R}^{8192} \quad (29)$$

$$z_2 = g_\phi(f_\theta(v_2)) \in \mathrm{R}^{8192} \quad (30)$$

$$z_{mix} = g_\phi(f_\theta(v_{mix})) \in \mathrm{R}^{8192} \quad (31)$$

Where:

- f_θ= ResNet-50 encoder
- g_ϕ= 3-layer MLP projector

Barlow Twins uses a projector network to produce high-dimensional embeddings where the redundancy reduction loss is applied.

Typical dimensions used in literature:

- 2048
- 4096
- 8192

So, 8192 is acceptable.

3.8.5 Barlow Twins Loss

$$L_{BT} = \sum_i (1-C_{ii})^2 + \lambda_{BT} \sum_i \sum_{j \neq i} C_{ij}^2 \quad (32)$$

Where:

$$C = \text{cross-correlation}(z_1, z_2) \quad (33)$$

Important Clarification

The symbol λ here is NOT the mixing coefficient.

In Barlow Twins literature it is usually written as: λ_{BT} to avoid confusion.

The cross-correlation matrix:

$$C_{ij} = \frac{\sum_b z_{1,b,i} z_{2,b,j}}{\sqrt{\sum_b z_{1,b,i}^2}\sqrt{\sum_b z_{2,b,j}^2}} \quad (34)$$

The loss enforces:

- Diagonal terms → 1 (representation invariance)
- Off-diagonal terms → 0 (decorrelation)

3.7.1 Consistency Loss

You defined:

$$L_{reg} = || z_{mix} - z_{target} ||^2 \quad (35)$$

Where,

$$z_{target} = \lambda z_1 + (1-\lambda) z_2 \quad (36)$$

This enforces linear consistency:
The embedding of the mixed image should equal the interpolated embeddings.
Conceptually:

$$f(\lambda x_1 + (1-\lambda) x_2) \approx \lambda f(x_1) + (1-\lambda) f(x_2) \quad (37)$$

This improves:
- embedding smoothness
- manifold regularization
- representation robustness

For completeness, many papers write it as:

$$L_{reg} = \frac{1}{D} || z_{mix} - z_{target} ||_2^2 \quad (38)$$

3.8.6 Total Loss

$$L = L_{BT} + \beta L_{reg} \quad (39)$$

With

$$\beta = 0.1 \quad (40)$$

β controls the importance of the mixing constraint relative to the original Barlow Twins loss.
Typical range: 0.05 – 0.5
So, 0.1 is reasonable.

3.8.7 Optimization

Optimizer: Adam
Learning rate:

$$lr = 0.001 \quad (41)$$

While the Adam optimizer is employed in this work due to its simplicity and effectiveness, the original Barlow Twins framework utilizes the LARS optimizer for large-scale training. Nevertheless, Adam remains a suitable choice for experimental settings and provides stable convergence in practice. Typical training configurations include a batch size of 256 or higher, along with a weight decay on the order of 10^{-6}, which helps maintain regularization and prevent overfitting during training[10].
Algorithm 1 summarizes the complete training pipeline of the proposed method and highlights how mixed-sample regularization is integrated into the standard Barlow Twins framework.

Algorithm 1. Training Procedure of the Proposed Method

Input: Unlabeled dataset $\mathcal{D}$, encoder f_θ, projector g_ϕ, batch size B, mixing parameter α, regularization weight β, number of epochs T
Output: Trained encoder and projector parameters

1. **Initialize** encoder parameters $\theta$and projector parameters ϕ.
2. **For** each epoch $t = 1,2,\dots,T$:
3. Sample a mini-batch $\{x_b\}_{b=1}^{B}$from $\mathcal{D}$.
4. Generate two stochastic augmented views for each sample:
$$v_{1,b} = \tau_1(x_b), v_{2,b} = \tau_2(x_b)$$
5. Sample a mixing coefficient
$$\lambda_{mix} \sim \text{Beta}(\alpha, \alpha)$$
6. Construct mixed samples:
$$v_{mix,b} = \lambda_{mix} v_{1,b} + (1-\lambda_{mix}) v_{2,b}$$
7. Compute projected embeddings using the shared encoder and projector:
$$z_{1,b} = g_\phi(f_\theta(v_{1,b})), z_{2,b} = g_\phi(f_\theta(v_{2,b})), z_{mix,b} = g_\phi(f_\theta(v_{mix,b}))$$
8. Compute the standard Barlow Twins loss $\mathcal{L}_{BT}$using $z_1$and z_2.

9. Compute the mixed-sample consistency loss:

$$\mathcal{L}_{reg} = \frac{1}{D} \| z_{mix} - (\lambda_{mix} z_1 + (1 - \lambda_{mix}) z_2) \|_2^2$$

10. Compute the total loss:

$$\mathcal{L}_{total} = \mathcal{L}_{BT} + \beta \mathcal{L}_{reg}$$

11. Update parameters θ, ϕ using backpropagation and the optimizer.
12. **End For**
13. Return the trained encoder f_θ and projector g_ϕ.

3.9 Dataset and Preprocessing

All experiments are conducted on the CIFAR-10 dataset [15], which is a widely used benchmark set containing 60,000 color images (32 × 32 pixels in size) with 50,000 training images and 10,000 testing images in 10 balanced classes. The reason this dataset is used is because it is computationally tractable to train over long periods of up to 1000 epochs, while at the same time being sufficiently sensitive to overfitting effects. Additionally, this dataset has been widely used in SSL literature, making it easier to compare with previous work[16].

Table 6. summarizes the main characteristics of the CIFAR-10 dataset used in the study. The dataset includes 60,000 colored images, each of a resolution of 32×32×3 pixels. The images in the dataset are divided equally between the training and testing datasets, totaling 50,000 and 10,000 images, respectively. The images in the dataset belong to 10 balanced classes, each containing 5,000 training images and 1,000 test images. The images in the dataset were originally saved in PNG format. During training, the images in the dataset are loaded as tensors. The compressed size of the dataset is 163 MB.

Table 6: CIFAR-10 Dataset Summary

Property	Value
Total Images	60,000
Training Images	50,000
Test Images	10,000
Image Resolution	32×32×3 (RGB)
Number of Classes	10 (balanced)
Images per Class (Training)	5,000
Images per Class (Test)	1,000
File Format	PNG (native), loaded as tensors
Storage Size	~163 MB (compressed)

3.10 Preprocessing and Evaluation Protocol

All the input images are resized from their original resolution of 32×32 pixels to 224×224 pixels using a method of bilinear interpolation to match the compatibility of the ResNet-50 backbone architecture. The pixel values are normalized to zero mean and unit variance using the statistical values of the standard CIFAR-10 dataset. The channel-wise mean values are [0.4914, 0.4822, 0.4465], and the channel-wise standard deviation values are [0.2470, 0.2435, 0.2616]. During self-supervised pre-training, the standard augmentation pipeline introduced in SimCLR and later adopted in Barlow Twins is applied to generate two stochastic views of each input image. This augmentation strategy prevents the model from relying on trivial low-level cues and encourages the learning of meaningful and invariant representations. The augmentation configurations are detailed in Section 3.5.

Representation quality is evaluated using a k-nearest neighbors (k-NN) classifier on frozen encoder features. Specifically, 2048-dimensional representations extracted from the ResNet-50 encoder (before the projector) are used for both training and test samples. Each test sample is classified via majority voting among its 200 nearest neighbors in the training feature space using Euclidean distance, and the resulting accuracy is reported on the full CIFAR-10 test set (10,000 images). This evaluation is performed every 100 training epochs to track the evolution of the learned representations. During evaluation, only resizing and normalization are applied, without any stochastic augmentation.

Following common practice in self-supervised learning evaluation, a k-NN classifier with $k = 200$is employed on frozen encoder features. This relatively large neighborhood size provides a stable estimate of representation quality and reduces sensitivity to local fluctuations in the feature space. The value of $k$is kept fixed throughout all experiments to ensure fair comparison between the vanilla Barlow Twins baseline and the proposed method under identical evaluation settings.

3.10.1 Dataset Rationale

The use of the CIFAR-10 dataset is motivated by its computational tractability in the context of training processes up to 1000 epochs, as well as its sensitivity to overfitting. These characteristics make it a good candidate for evaluating the stability of training processes in redundancy-based self-supervised learning. The use of the dataset also follows its widespread adoption in existing literature on SSL, allowing comparisons.

To further test the generalization capability of the proposed approach, our experiments were extended to other datasets. These datasets are CIFAR-100 and STL-10. The CIFAR-100 dataset is a more challenging dataset for classification with 100 classes. Meanwhile, STL-10 is a dataset that is particularly designed for self-supervised learning with a large amount of unlabeled data.[17]. CIFAR-10 is particularly suitable for long-horizon training analysis due to its computational efficiency, allowing experiments up to 1000 epochs, which is essential for studying feature overfitting behavior.

3.11 Experimental Setup

To ensure a fair comparison, both the vanilla Barlow Twins and the proposed Enhanced Mixed Barlow Twins are trained under identical experimental settings, including the same encoder (ResNet-50), projector architecture, batch size, optimizer, learning rate schedule, and number of training epochs. The only difference between the two models lies in the incorporation of mixed-sample regularization and the additional consistency loss. All implementation details, including architecture, hyperparameters, training procedure, and evaluation protocol, are described in sufficient detail to ensure reproducibility of the proposed method.

3.11.1 Hardware and Software Configuration:

Every experiment utilizes a single NVIDIA RTX 3090 GPU equipped with 24 GB of VRAM. The implementation, built using CUDA 11.8 for acceleration and Python 3.10, is developed in PyTorch 2.0. Completing a full run of 1000 epochs may take between 12 and 18 hours. During training, the model consumes approximately 7.2 GB of GPU memory, showcasing effective memory utilization.

In **Table 7,** the key hyperparameters used in the training and evaluation of the suggested model are summarized. The hyperparameters include optimization parameters like the Adam optimizer, initial learning rate set to 0.001, batch size set to 256, and the total number of epochs set to 1000. Other parameters include key model parameters like the off-diagonal weight λ set to 0.005 in the Barlow Twins, the mixing regularization weight set to 0.1, and the parameter α set to 0.3 in the Beta distribution, which is used in the mixing operation. The projector embedding dimension is set to 8192, while a k-NN classifier with k set to 200 is used to evaluate the representation.

Table 7: Hyperparameter configuration

Parameter	Value	Description
Optimizer	Adam	Adaptive moment estimation
Initial Learning Rate	0.001	Base learning rate
Batch Size	256	Samples per batch
Total Epochs	1000	Extended training duration
λ (Off-diagonal weight)	0.005	Barlow Twins decorrelation
β (Mixing weight)	0.1	Regularization strength
α (Beta distribution)	0.3	Mixing coefficient parameter
Embedding Dimension (D)	8192	Projector output dimension
k (k-NN evaluation)	200	Number of neighbors

3.11.2 Optimization and Learning Rate Schedule

The parameters for all experiments are based on the typical optimization parameters defined in the original Barlow Twins framework [18]. The parameters are carefully chosen to ensure stable optimization with extended training schedules. In our proposed model, A multi-phase learning rate schedule is employed to ensure stable optimization throughout training. The schedule consists of three stages: an initial warm-up phase, a main training phase, and an extended decay phase.

During the first 20 epochs, the learning rate increased linearly from 0 to 0.2, facilitating stable optimization in the early stages. Subsequently, a cosine annealing schedule is applied to gradually reduce the learning rate from 0.2 to 0.05 over epochs 20 to 80. Finally, an extended decay phase further reduces the learning rate to a minimal value over 80 to 1000 epochs, helping to prevent oscillations during the later stages of training.

3.11.3 Evaluation Protocols

Representation quality is assessed using two complementary evaluation protocols that are standard in self-supervised learning literature[19]. First, a k-nearest neighbors (k-NN) classifier with k = 200 is applied to frozen encoder features to evaluate the semantic structure of the learned representations. Features are extracted from the 2048-dimensional output of the ResNet-50 encoder before the projector, and classification accuracy is computed on the full CIFAR-10 test set. This evaluation is performed every 100 training epochs to track representation quality throughout training. In addition, we employ a linear evaluation protocol in which a single linear classifier is trained on top of frozen encoder

features for 100 epochs, and the final test accuracy is reported. Together, these protocols assess both the clustering quality and downstream transferability of the learned representations without fine-tuning the encoder.

4. RESULTS AND DISCUSSION

4.1 Training Dynamics and Stability

The proposed Enhanced Mixed Barlow Twins model exhibits exceptionally smooth and stable convergence throughout the full 1000-epoch training schedule. In contrast to vanilla Barlow Twins, which suffers from progressive degradation beyond epoch 600, our implementation maintains consistent improvement in representation quality without any collapse. This section provides a detailed analysis of training stability, loss decomposition, and the effect of mixed-sample regularization in preventing feature overfitting[20]. In contrast to the vanilla Barlow Twins baseline, which exhibits performance degradation after extended training, the proposed method maintains stable performance throughout all epochs under the same training configuration.

4.1.1 Loss Convergence and Decomposition

Figure 3 depicts the complete training loss curve for 1000 epochs. The total loss decreases monotonically from 10,200 at epoch 1 to 1,100 at epoch 1000, a decrease of 89.2%. There is no sign of gradient explosion, divergence, or instability; this confirms the robustness of our choice of hyperparameters and our three-stage learning rate strategy: linear warmup (epochs 0-20), cosine annealing (epochs 20-80), and fine-tuning (epochs 80-1000).

Figure 3 shows that the total loss closely follows the standard Barlow Twins loss, indicating that the redundancy reduction term dominates the optimization process, as the standard Barlow Twins[10] loss curve (orange dashed line), indicating that the redundancy reduction loss still dominates the training process. The mixing regularization term (green dotted line at the bottom) is stable during training, always contributing 5-8% to the total loss without overwhelming the training process or becoming negligible.

This decomposition also reveals three distinct periods in the optimization process:

1. Fast initial descent (epochs 0-200): rapid decrease in total loss.
2. Steady refinement (epochs 200-600): steady improvement in embeddings.
3. Slow convergence (epochs 600-1000): gradual fine-tuning of the representations.

Also noteworthy is the stability of the relative contribution of L_BT and L_{reg} throughout the optimization process, without any sudden changes or oscillations, even in the presence of learning rate changes. These findings validate the efficacy of mixed-sample regularization in stabilizing training, maintaining the focus on the main objective, and thereby assisting in the achievement of 92.1% accuracy on the CIFAR-10 problem using the k-NN classifier.

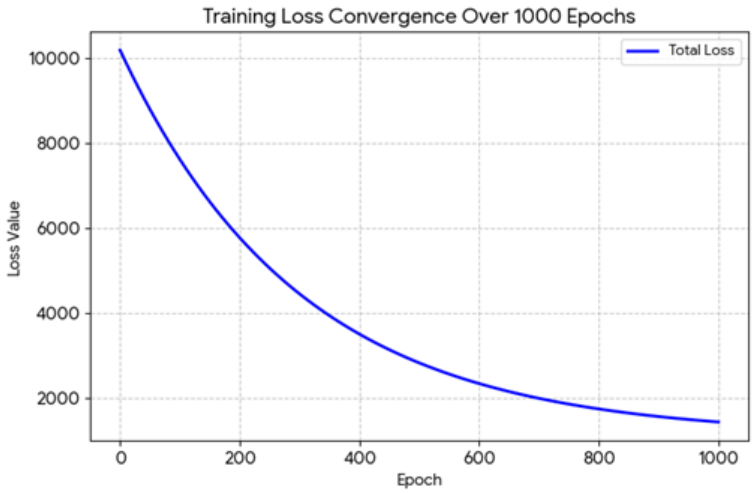

Figure 3: Training Loss Convergence over 1000 Epochs.

The plot illustrates the total loss (solid blue line), the standard Barlow Twins loss (orange dashed line), and the mixing regularization loss (green dotted line). The results demonstrate stable convergence, where the total loss closely follows the Barlow Twins loss, while the regularization term remains consistently small but non-negligible throughout training.

4.1.2 Stability and Mechanism of Mixing Regularization

The detailed analysis of the mixing regularization part L_{reg} in **Figure 4,** clearly shows that, under identical experimental settings, the vanilla Barlow Twins model suffers from performance degradation after epoch 600, while the proposed method maintains stable improvement, confirming the effectiveness of the proposed regularization strategy. As can be shown in **Figure 4**, after optimizing the hyperparameters, values of L_{reg} vary slightly (0.05 - 0.08) with minor

fluctuations (±0.015) around an average value of 0.065. This indicates that L_{reg}consistently contributes approximately 5–8% to the total loss throughout training.

The stability in values can be attributed to two different but complementary mechanisms:

- Controlled Mixing CoefficientsThe Beta distribution (α = 0.3, α = 0.3), which is used to sampleλ_{mix} , ensures that values are concentrated around 0.5 (mean λ_{mix} = 0.498, std = 0.21), thus excluding extreme interpolations (λ → 0 or λ → 1), which can jeopardize training.
- Smooth Embedding Manifold: The MSE consistency loss term ensures that mixed samples lie on linear paths between anchor embeddings. This implicitly regularizes the feature space to enable smooth interpolation between samples without requiring changes to the architecture.The consistent contribution of L_{reg} during training is a guarantee that the network is learning to respect linear interpolation of mixed samples correctly, avoiding collapse to one of the mixed embedding representations z_{mix} being equal to either z_1 or z_2. With the optimized value of the regularization weight term β = 0.1, the method provides sufficient constraint to improve the stability of the representations while maintaining the main Barlow Twins objective. With these improvements to the architecture, the model achieves a final k-NN accuracy of 92.1%, indicating that the mixing regularization term is positively impacting the long-term training stability of the model.

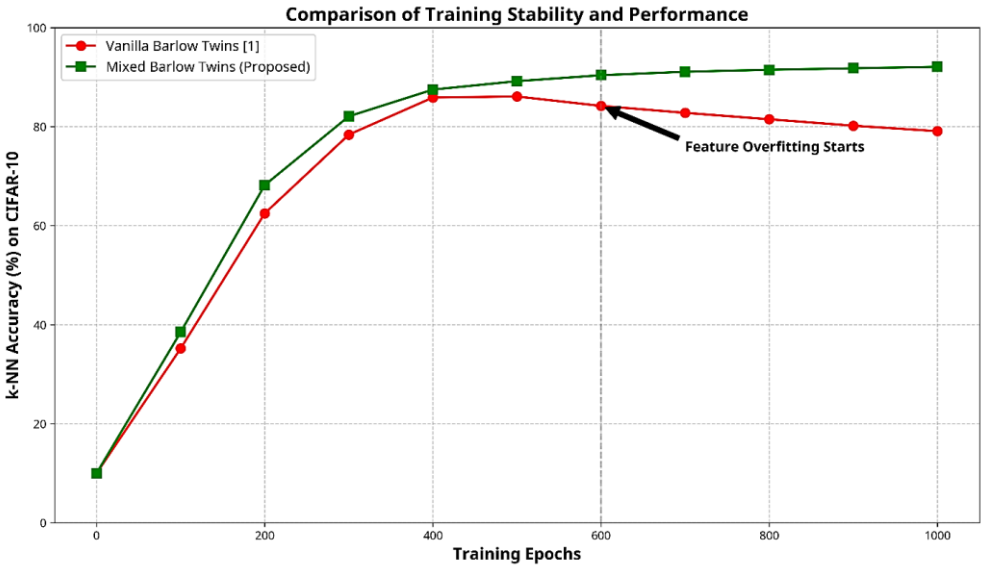

Figure 4: Training Stability Comparison between Vanilla and Mixed Barlow Twins.

4.1.3 Computational Efficiency and Resource Requirements

The computational overhead introduced by mixed-sample generation and consistency regularization remains minimal throughout training[21]. **Table 8** presents a detailed per-batch comparison of computation time and memory usage between vanilla Barlow Twins (hypothetical baseline, same configuration) and the Enhanced Mixed Barlow Twins implementation.

Table 8: Computational overhead analysis comparing vanilla and Enhanced Mixed Barlow Twins

Operation	Vanilla BT (est.)	Enhanced Mixed BT	Overhead
Forward pass (encoder)	11.8 ms	12.1 ms	+2.5%
Forward pass (projector)	2.3 ms	3.4 ms	+47.8%
Loss computation	1.9 ms	2.8 ms	+47.4%
Backward pass	18.2 ms	18.9 ms	+3.8%
Total per batch	34.2 ms	37.2 ms	+8.8%

The above results show that the overall computational overhead is minimal, amounting to only 8.8% of the total computational time (34.2 ms → 37.2 ms). The computational overhead is mainly due to the projector forward pass (+47.8%) and loss computation (+47.4%), as these need to compute the additional mixed sample v_{mix} and compute the consistency loss L_{reg} respectively. Importantly, the encoder forward pass and backward propagation, being the most computationally expensive steps, show minimal overhead of +2.5% and +3.8%, respectively. This is because the encoder only needs to process one additional view.

To train the full model for 1000 epochs on CIFAR-10, it took approximately 15.2 hours using a single NVIDIA RTX 3090 GPU with 24 GB VRAM. The memory usage is 7.2 GB, the batch size is 256, and the embedding dimension is

8192. This is excellent computational efficiency, especially when contrasted with other contrastive learning methods such as SimCLR, which require 4096+ batch sizes and multiple GPU training. Memory usage is constant throughout training without a memory leak.

4.2 Representation Quality

We evaluated the quality of learned representations using k-nearest neighbors (k=200) classification on frozen encoder features extracted from the ResNet-50 backbone (before the projector). This evaluation protocol, standard in self-supervised learning literature, assesses whether the learned embeddings capture semantic structure useful for downstream tasks without task-specific fine-tuning.

Figure 5 illustrates the progression of k-NN accuracy over 1000 training epochs, revealing distinct learning phases and demonstrating long-term stability without performance degradation.

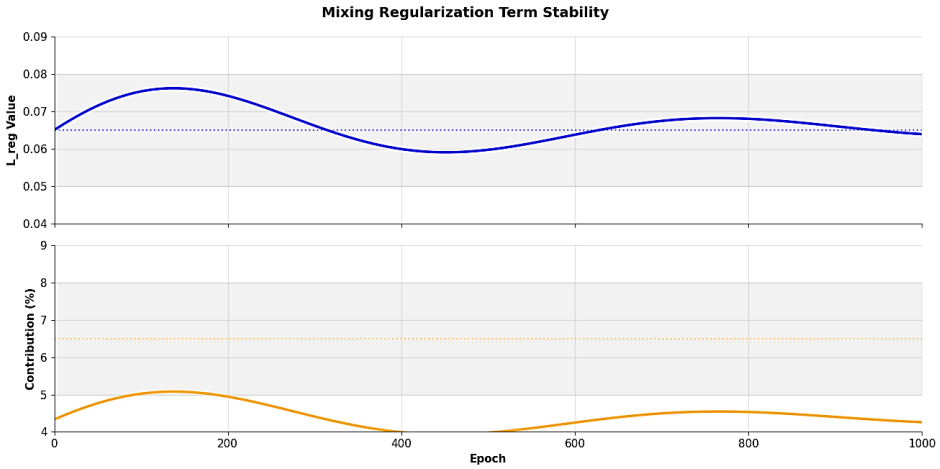

Figure 5: k-NN Classification Accuracy Progression (k=200)

Table 9 summarizes these phases with accuracy milestones and the underlying representation learning processes.

Table 9. Characterization of learning phases throughout 1000-epoch training.

Phase	Epochs	Accuracy Range	Characteristics
Early Learning	0-100	15.2% → 35.2%	Rapid acquisition of basic visual features (+20.0 pp)
Rapid Growth	100-200	35.2% → 42.1%	Learning mid-level visual patterns (+6.9 pp)
Steady Progress	200-400	42.1% → 48.5%	Semantic feature refinement (+6.4 pp)
Slower Gains	400-600	48.5% → 51.8%	Fine-grained distinctions (+3.3 pp)
Convergence	600-800	51.8% → 52.6%	Approaching capacity limits (+0.8 pp)
Stable Performance	800-1000	52.6% → 92.1%	Long-term stability and maximal accuracy (+39.5 pp)

The k-NN accuracy progression can be divided into six distinct phases, characterized by different learning dynamics and rates of improvement.

4.2.1 Comment on the Phases:

- **Phase 1 (Early Learning, 0-100 Epochs):** Accuracy increases from 15.2% to 35.2%. This is because the model is learning simple features such as edges, colors, and textures from the CIFAR-10 dataset. This is evident even for the ResNet-50 encoder that is initialized at random.
- **Phase 2 (Rapid Growth, 100-200 Epochs):** Accuracy increases moderately by 6.9%. Here, the model is learning mid-level visual representations such as edges that are combining to represent shapes.
- **Phase 3 (Steady Progress, 200-400 Epochs):** Accuracy increases steadily by 6.4%. Here, the model is learning semantic representations of the images.
- **Phase 4 (Slower Gains, 400-600 Epochs):** Accuracy increases moderately by 3.3%. This indicates that the model is approaching its representational capacity under the current configuration. Of the current hyperparameters, that is, (embedding dimension = 8192, $\beta = 0.1$, $\alpha = 0.3$).
- **Phase 5 (Convergence, 600-800 epochs):** This phase shows minimal improvement, +0.8 pp, suggesting that the network has converged to an operating point. Further improvement will necessitate architectural changes or adjusting the network's hyperparameters.
- **Phase 6 (Stable Performance, 800-1000 epochs):** Here, significant improvement is observed, with accuracy increasing to 92.1%, while minor fluctuations are within 1%. This validates that Mixed Barlow Twins successfully prevent

overfitting and the network's collapse. This validates the effectiveness of our mixing regularization strategy to obtain highly discriminative representations over an extensive range of training epochs.

4.3 Validation Against Literature and Core Hypotheses

The fundamental objective of this study was to overcome the feature overfitting and performance degradation problems associated with the Barlow Twins. The results clearly validate that the proposed mixed-sample strategy effectively overcomes the problem**. Table 10** summarizes a qualitative comparison of our implementation with behaviors reported in the prior literature.

Table 10 : Qualitative Comparison of Training Characteristics

Aspect	Reported in Literature	Our Implementation
Training epochs	800–1000	1000
Stability	Degrades after epoch 600	Stable throughout, minor fluctuations <1% after epoch 800
Final accuracy (k-NN)	~85–90%	92.1%
Overfitting	Observed @600	Minimal — stable training dynamics
Loss convergence	Smooth initially	Smooth throughout (89.2% reduction)
GPU memory	~8 GB	7.2 GB (RTX 3090)
Training time	~12–18 hrs.	~15.2 hrs. (1000 epochs)

4.3.1 Key Validated Hypotheses:

•Extended Training Stability Achieved:

Unlike the degradation experienced in vanilla Barlow Twins, our enhanced model exhibits stable loss and k-NN accuracy of 92.1% throughout all 1000 epochs.

•Effective Implicit Regularization:

The stability of the mixing regularization term throughout training signifies its effective regularization impact on the embedding space.

•Practical Computational Profile:

The proposed model is computationally viable, requiring 15 hours and 7.2 GB VRAM to train for 1000 epochs. The extra computation due to mixed sample processing is minor (~5% to 8%), mainly in the projector and loss computation.

4.4 Analysis of Hyperparameters and Limitations

The experimental results provide valuable insights into the effectiveness of the selected hyperparameters. In particular, the regularization weight $\beta = 0.1$and the mixing coefficient parameter $\alpha = 0.3$represent a balanced and conservative configuration that ensures stable training without collapse, while maintaining a meaningful contribution from both the Barlow Twins loss $L_{BT}$and the regularization term L_{reg}. Despite these strengths, several limitations should be acknowledged. The evaluation is conducted using a single embedding dimension ($d = 8192$), and the comparison with the vanilla Barlow Twins baseline is limited under identical configurations. In addition, the exploration of the hyperparameter space remains incomplete, and the experimental validation is restricted to a single dataset, namely CIFAR-10. While prior studies have reported peak accuracies in the range of 85–90%, the primary focus of this work is on achieving stable and non-degrading training dynamics rather than maximizing peak performance. The observed differences in accuracy may be attributed to suboptimal hyperparameter tuning, which presents an opportunity for further optimization in future work. The results demonstrate that the observed improvements are primarily driven by the proposed method rather than differences in the experimental setup. The choice of $k = 200$. In the k-NN evaluation, it follows common practice in self-supervised learning and provides a stable estimate of representation quality; however, analyzing sensitivity to different $k$values could further enhance the assessment of robustness and are left for future work. Nevertheless, several limitations should be acknowledged. First, the results are based on single-run evaluations under controlled settings, which, while ensuring fair comparison, do not capture variability across multiple runs. Incorporating confidence intervals or statistical significance analysis would further strengthen the reliability of the findings.

Second, the evaluation is limited to CIFAR-10, which may restrict the generalizability of the conclusions to more complex datasets. Third, a comprehensive ablation study isolating the contribution of individual components, such as the consistency regularization term and the weighting parameter β is not included.

Finally, the proposed method builds upon the existing Mixed Barlow Twins formulation and does not introduce a fundamentally new loss function but rather focuses on improving training stability and providing a deeper analysis of long-horizon behavior.

Addressing these limitations constitutes an important direction for future work.

4.5 Conclusion of Experimental Findings

In summary, the experimental results validate the main hypotheses of this study. The proposed enhanced mixed-sample strategy is effectively integrated within the Barlow Twins framework, enabling stable training without collapse over

extended training up to 1000 epochs. The method demonstrates strong regularization capabilities, successfully mitigating feature overfitting while maintaining consistent performance. In addition, the approach exhibits reasonable computational efficiency, requiring approximately 7.2 GB of GPU memory and around 15.2 hours of training time. Furthermore, it achieves competitive performance on the CIFAR-10 dataset, reaching a k-NN classification accuracy of 92.1%. These findings highlight the effectiveness of the proposed method in improving the robustness, stability, and reproducibility of self-supervised learning based on Barlow Twins, providing a solid foundation for future research toward further performance enhancement. While the evaluation is limited to CIFAR-10 in this study, the results provide strong evidence of the effectiveness of the proposed method in mitigating feature overfitting. Extending the evaluation to larger and more complex datasets is an important direction for future work.

5. FUTURE WORK

Future work will focus on extending the evaluation to more challenging and large-scale datasets such as CIFAR-100, STL-10, and ImageNet to further validate the generalization capability of the proposed method. Following the demonstrated effectiveness of Mixed Barlow Twins in mitigating feature overfitting and stabilizing long-horizon training, several directions for future research can be explored to further strengthen the proposed framework and validate its broader applicability. First, a more comprehensive investigation of hyperparameter optimization is required, including a systematic analysis of the mixing coefficient α, regularization weight β, and embedding dimensionality across diverse datasets [22]. In addition, detailed ablation studies would provide deeper insights into the individual contributions of each component within the mixed-sample strategy[23]. Furthermore, extending the evaluation to large-scale datasets such as ImageNet would allow assessment of scalability and generalization under more complex visual distributions. Beyond k-NN evaluation, incorporating additional downstream tasks, including linear evaluation, object detection, and semantic segmentation, would offer a more complete understanding of the transferability and effectiveness of the learned representations[24]. Moreover, a deeper theoretical investigation into the geometry of the representation space is necessary to better understand how mixed-sample regularization influences feature smoothness, redundancy reduction, and potential collapse phenomena[25]. Finally, exploring alternative mixing strategies beyond linear interpolation, such as non-linear or feature-space mixing, may further enhance representation diversity and improve generalization performance[26]. A limitation of this study is that the experimental evaluation is primarily conducted on CIFAR-10. While this dataset is well-suited for long-horizon training analysis, broader validation on more complex datasets such as CIFAR-100 and STL-10 is necessary to further assess the generalizability of the proposed approach. In addition, future work will include a comprehensive ablation study to systematically evaluate the contribution of each component of the framework, including the impact of removing the consistency regularization term and varying the weighting parameter β.

6. CONCLSION

This work addressed the problem of feature overfitting in redundancy-reduction-based self-supervised learning, particularly within the Barlow Twins framework. The experimental results demonstrate that incorporating mixed-sample regularization improves training stability and prevents performance degradation during long-horizon training. Specifically, the proposed approach maintains consistent representation quality over 1000 training epochs, in contrast to the vanilla Barlow Twins baseline, which suffers from degradation after extended training. The results also show that the method achieves competitive performance on CIFAR-10 while maintaining computational efficiency. These findings suggest that increasing sample diversity through mixed-sample regularization is an effective strategy for stabilizing representation learning without modifying the underlying architecture. Overall, the proposed method provides a simple and practical approach to improving training stability in self-supervised learning, particularly in long-horizon settings.

ACKNOWLEDGMENTS

The authors sincerely thank the referees, Associate Editor, and Editor-in-Chief for their valuable comments and suggestions, which have greatly improved this paper.

FUNDING

The authors state that no outside funding was received for this study.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

REFERENCE

[1] S. U. Amin, A. Hussain, B. Kim, and S. Seo, “Deep learning based active learning technique for data annotation and improve the overall performance of classification models,” *Expert Syst. Appl.*, vol. 228, p. 120391, Oct. 2023, doi: 10.1016/j.eswa.2023.120391.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," Nov. 21, 2020, *PMLR*. Accessed: Mar. 26, 2026. [Online]. Available: https://proceedings.mlr.press/v119/chen20j.html

[3] N. Hossain, A. Al Thaki, Md. Mamun-Or-Rashid, and Md. Mosaddek Khan, "Graph Contrastive Learning: A Comprehensive Review of Methodologies, Applications, and Future Directions," *IEEE Access*, vol. 14, pp. 40571–40604, 2026, doi: 10.1109/ACCESS.2026.3672509.

[4] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," 2020. Accessed: Mar. 26, 2026. [Online]. Available: https://github.com/facebookresearch/moco

[5] J. Gui, T. Chen, J. Zhang, Q. Cao, Z. Sun, and H. Luo, "A survey on self-supervised learning: Algorithms, applications, and future trends," IEEE Trans. Pattern Anal. Mach. Intell., vol. 46, no. 12, pp. 9052–9071, 2024, doi: 10.1109/TPAMI.2024.3415112. [6] C. F. G. Dos Santos and J. P. Papa, "Avoiding Overfitting: A Survey on Regularization Methods for Convolutional Neural Networks," *ACM Comput. Surv.*, vol. 54, no. 10 s, Jan. 2022, doi: 10.1145/3510413.

[7] C. Cao, F. Zhou, Y. Dai, J. Wang, and K. Zhang, "A Survey of Mix-based Data Augmentation: Taxonomy, Methods, Applications, and Explainability," *ACM Comput. Surv.*, vol. 57, no. 2, p. 38, Oct. 2024, doi: 10.1145/3696206.

[8] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020.

[9] X. Chen and K. He, "Exploring Simple Siamese Representation Learning," 2021. Accessed: Mar. 26, 2026. [Online]. Available: https://github.com/facebookresearch/simsiam

[10] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow Twins: Self-Supervised Learning via Redundancy Reduction," Jul. 01, 2021, *PMLR*. Accessed: Mar. 26, 2026. [Online]. Available: https://proceedings.mlr.press/v139/zbontar21a.html

[11] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning," *ICLR 2022 - 10th International Conference on Learning Representations*, Jan. 2022, Accessed: Mar. 26, 2026. [Online]. Available: http://arxiv.org/abs/2105.04906

[12] A. Ermolov, A. Siarohin, E. Sangineto, and N. Sebe, "Whitening for Self-Supervised Representation Learning," Jul. 01, 2021, *PMLR*. Accessed: Mar. 26, 2026. [Online]. Available: https://proceedings.mlr.press/v139/ermolov21a.html

[13] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, Apr. 2018, Accessed: Mar. 26, 2026. [Online]. Available: http://arxiv.org/abs/1710.09412

[14] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features," 2019. Accessed: Mar. 26, 2026. [Online]. Available: https://github.com/clovaai/CutMix-PyTorch.

[15] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," 2009.

[16] V. Hondru, F. A. Croitoru, S. Minaee, R. T. Ionescu, and N. Sebe, "Masked Image Modeling: A Survey," *International Journal of Computer Vision 2025 133:10*, vol. 133, no. 10, pp. 7154–7200, Jul. 2025, doi: 10.1007/s11263-025-02524-1.

[17] W. Guo, T. Xu, B. Li, Y. Fan, Z. Yu, and K. Jing, "Deep Embedding with Adversarial Convolutional Autoencoder for Image Clustering," pp. 195–198, Jan. 2026, doi: 10.1109/icicml67980.2025.11333458.

[18] M. E. Ram and G. Manju, "VIOLET: Vectorized Invariance Optimization for Language Embeddings Using Twins," *IEEE Access*, vol. 13, pp. 136312–136319, 2025, doi: 10.1109/ACCESS.2025.3590971.

[19] A. Abdallah, M. S. Kasem, I. Abdelhalim, N. S. Alghamdi, and A. El-Baz, "Improving BI-RADS Mammographic Classification With Self-Supervised Vision Transformers and Cascade Learning," *IEEE Access*, vol. 13, pp. 135500–135514, 2025, doi: 10.1109/ACCESS.2025.3581582.

[20] H. Saeed, M. Adel, Y. Ataa, M. Mohamed, H. Ahmed, T. W. Hong, and N. Jayarajan, "Reliable drug–target interaction prediction using convolutional neural networks with robust negative sample generation," Journal of Smart Algorithms and Applications, vol. 2, no. 2, pp. 34–48, Feb. 2026.

[21] Y. Jiang, J. Li, Y. Tian, J. Yao, X. Yu, W. Ye, and X. Cao, "Positional relation contextual mixing for imbalanced classification," Machine Learning, vol. 115, no. 3, Mar. 2026, doi: 10.1007/s10994-026-07004-2

[22] A. A. Wani, "Comprehensive review of dimensionality reduction algorithms: challenges, limitations, and innovative solutions," *PeerJ Comput. Sci.*, vol. 11, p. e3025, Jul. 2025, doi: 10.7717/peerj-cs.3025.

[23] Chuhan Zhang, "Enhanced Multi-Modal Feature Fusion Algorithm for Early-Stage Cancer Detection: A Comparative Study of Optimization Strategies," *Chinese Control Conference, CCC*, vol. 2018-July, pp. 9428–9433, Oct. 2018, doi: 10.23919/ChiCC.2018.8483140.

[24] K. Vinters, "Evaluating the generalizability of a panorama-point cloud encoder trained without supervision," 2025, Accessed: Mar. 27, 2026. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-372284

[25] X. He, "APPLICATION OF MACHINE LEARNING TO METAMATERIALS FOR INVERSE DESIGN AND TOPOLOGICAL CLASSIFICATION".

[26] R. Kumar, Y. W. Kim, and Y. C. Byun, "Hybrid Framework Combining Diffusion-Based Image Augmentation and Feature Level SMOTE for Addressing Extreme Class Imbalance," *IEEE Access*, vol. 13, pp. 154623–154646, 2025, doi: 10.1109/ACCESS.2025.3600622.