International Scientific Technical and Economic Research **(ISSN: 2959-1309)**
Vol. 4, No. 2, (June 2026), pp. 78-97
Journal homepage: https://www.istaer.online

RESEARCH ARTICLE **DOI: 10.71451/ISTAER2616**

Research on Image Representation Learning Method Based on Self-Supervised Learning

Juanpeng Zhang*

Department of Electrical Engineering, Cheongju University, Cheongju, Seoul, Republic of Korea

Abstract: Aiming at the problems of negative sample dependence, representation degradation, and insufficient cross-scale modeling in self-supervised image representation learning, this paper proposes a self-supervised learning framework that combines multi-view consistent learning and cross-scale feature fusion. This method constructs a multi-branch collaborative structure, introduces a non-negative sample optimization strategy and a feature distribution constraint mechanism, and achieves efficient mining and stable expression of image semantic information. On the ImageNet dataset, the accuracy of linear evaluation reached 77.8%, which was 8.5% and 2.5% higher than that of SimCLR and SwAV, respectively; In downstream tasks, the target detection mAP increased by about 2.5%, and the semantic segmentation mIoU increased by about 2.5%. At the same time, the accuracy improves by 7.5% under noise disturbance, demonstrating stronger robustness. The experimental results show that this method is superior to the existing mainstream methods in terms of characterization quality, generalization ability and training stability, and has good application potential.

Keywords: Self-supervised learning; Image representation learning; Cross-scale feature fusion; Non-negative sample learning; Deep learning

How to Cite: Zhang, J. (2026). Research on Image Representation Learning Method Based on Self-Supervised Learning. *International Scientific Technical and Economic Research*, *4*(2), 78–97. https://doi.org/10.71451/ISTAER2616

Article history: Received: 20 Jan 2026; **Revised:** 27 Feb 2026; **Accepted:** 03 Apr 2026; **Published:** 12 Apr 2026

1. INTRODUCTION

In recent years, with the rapid development of deep learning, learning high-quality visual representations in the absence of large-scale annotated data has become an important research direction in computer vision [1],[2]. As a method that mines potential structural information from data without manual annotation, self-supervised learning has gradually shown great potential in image representation learning [3],[4],[5]. By constructing pre-training tasks or designing consistency constraints, self-supervised methods can guide the model to learn feature representations with semantic information, thereby achieving performance close to or even better than supervised learning on downstream tasks [6]. At present, self-supervised learning has been widely used in image classification, object detection, semantic segmentation, and other tasks, and has gradually become an important paradigm for visual pre-training. However, while

***Corresponding author:** Juanpeng Zhang, Department of Electrical Engineering, Cheongju University, Cheongju, Republic of Korea. Email: zjp6574@gmail.com

significant progress has been made, the field still faces a series of core challenges to be solved, which limits its further development and application.

From existing research, mainstream self-supervised methods mainly include three paradigms: contrastive learning, generative modeling, and clustering-based learning. By constructing positive and negative sample pairs, contrastive learning methods strengthen consistency among similar samples and widen the distance between different samples, achieving outstanding results in practice; Generative methods learn the latent structure of input data through image reconstruction or prediction tasks [7],[8]. Clustering methods transform unsupervised problems into classification problems by introducing pseudo-labels or prototype representations. Although these methods show certain advantages in different scenarios, they still face some common bottlenecks. First, many contrastive learning methods rely heavily on a large number of negative samples, and their performance often depends on large batch sizes or additional storage structures, which not only increases computational and storage costs but may also introduce semantic conflicts. Second, although non-negative sample methods alleviate the above problems to some extent, they are prone to representation degradation during training, i.e., the model output tends to become constant, thus losing discriminative ability. Third, existing methods are still insufficient in cross-scale semantic modeling. They often struggle to simultaneously account for local details and global structural information, resulting in insufficient feature expression [9]. In addition, most methods are sensitive to data augmentation strategies. Different augmentation methods may significantly affect model performance and reduce the stability and generalization ability of the methods.

The core idea is to enhance the model's ability to understand complex visual structures through collaborative modeling of multi-perspective and multi-scale information, and to introduce reasonable constraint mechanisms to improve the discriminability and stability of feature distributions. In the overall framework design, this paper attempts to combine a non-negative sample learning strategy with a structural asymmetry mechanism to avoid the negative sample dependence problem in traditional contrastive learning, and to further optimize representation quality through feature enhancement and distribution constraints. At the same time, by building a multi branch collaborative learning structure, information from different perspectives and levels can interact effectively under a unified framework, so as to achieve more comprehensive feature learning.

2. RELATED WORK

In recent years, self-supervised learning has made significant progress in the field of image representation learning, and the method based on comparative learning has become one of the mainstream research directions [10],[11],[12],[13]. Representative work, such as SimCLR, has achieved strong representation ability by constructing positive and negative sample pairs and training with the help of large-scale batch data; Moco introduces the momentum encoder and dynamic dictionary queue mechanism, which effectively alleviates the dependence on large-scale computing resources under the condition of small batch; InfoNCE loss, as the core optimization goal, is widely used in the comparative learning framework. By maximizing the positive sample similarity and minimizing the negative sample similarity, the feature space is promoted to form a discriminant structure. However, such methods generally rely on a large number of high-quality negative samples, and their performance is highly dependent on the number and diversity of negative samples [14],[15],[16]. At the same time, large batches or additional storage structures are required to maintain sample comparison, which brings high computational overhead and storage costs to a certain extent. In addition, the potential semantic conflict in negative samples may also have an adverse impact on model training, limiting its further development.

In order to get rid of the dependence on negative samples, self-monitoring methods without negative samples are gradually emerging, among which BYOL and SimSiam are representative works. This kind of method achieves feature alignment without explicitly using negative samples by constructing a double branch structure and introducing a predictor and a target network [17]. BYOL stabilizes the target network through momentum update mechanism to avoid model degradation, while SimSiam proves that it can obtain effective characterization without momentum encoder under certain conditions through structural simplification. Although these methods have advantages in computational efficiency and simplicity, their training process is sensitive to structural design and prone to collapse, that is, the model output tends to constant mapping, thus losing the ability of discrimination [18]. Therefore, how to ensure the stability of training under the framework of no negative samples is still an important research issue in this direction.

Another important research direction is the self-monitoring method based on clustering and prototype learning. For example, DeepCluster classifies the unlabeled data into pseudo categories through iterative clustering and feature learning, so as to guide the model to learn the semantic structure; SwAV further introduces online clustering and exchange prediction mechanism to achieve efficient training through multi view feature allocation [19],[20]. This kind of method avoids the dependence on explicit negative samples to a certain extent, and can capture the potential category structure in the data. However, because the clustering process itself depends on the quality of initial features, the pseudo tags generated by it may be unstable, leading to semantic drift or noise accumulation of categories, which will affect the convergence effect of the model. In addition, in complex scenes, when the category boundary is fuzzy or the data distribution is uneven, the performance of clustering methods still has some limitations.

With the deepening of research, multi perspective and multi-scale representation learning has gradually become an important means to improve the performance of the model [21],[22],[23]. By applying different data enhancement to the same image, the multi view method makes the model learning maintain the semantic consistent representation under different observation conditions, so as to improve the robustness; The multi-scale method enhances the model's perception of local details and global structure by fusing features of different levels or resolutions [24],[25]. Some researches combine pyramid structure or attention mechanism to achieve cross layer feature interaction and fusion. However, such methods often face the problem of low efficiency of feature fusion in practical applications, especially in the multi branch or multi-scale structure, the complex information transmission path easily leads to redundant computing and feature redundancy, which limits the further improvement of the overall performance.

Compared with the above methods, this paper makes systematic improvements from three aspects: structure design, learning strategies and optimization mechanism. In terms of structure, through the construction of multi branch collaborative framework and the introduction of cross scale feature fusion module, the efficient integration of different levels of information is achieved; In terms of the optimization mechanism, the comparative learning strategy without negative samples, combined with asymmetric structure and distribution constraints, effectively alleviates the representation collapse problem and improves the training stability [26],[27]. Compared with the existing methods, this method not only avoids the negative sample dependence, but also improves the quality of feature expression and training efficiency, so as to achieve better performance on multiple tasks and data sets.

3. METHODOLOGY

3.1 Overall frame design

This paper proposes a unified framework for image representation learning based on self-supervised learning, which consists of a feature encoder, a projection head, a predictor, and a memory module. Given the input image x, multiple views $\{v_i\}_{i=1}^{N}$ are generated through

Zhang., ISTAER. 2616., 12 Apr 2026 https://doi.org/10.71451/ISTAER2616

random data augmentation, and each view is fed into the encoder $f_\theta(\cdot)$ with shared parameters to obtain the basic feature representation [28],[29]:

$$h_i = f_\theta(v_i) \tag{1}$$

Where θ is the encoder parameter, and $h_i \in \mathbb{R}^d$ is the high-dimensional feature vector of the ith view angle. Then it is mapped to the contrast space through the projection head $g_\phi(\cdot)$:

$$z_i = g_\phi(h_i) \tag{2}$$

A predictor $q_\psi(\cdot)$ is used to construct an asymmetric structure to enhance training stability:

$$p_i = q_\psi(z_i) \tag{3}$$

The memory module is used to maintain the historical feature distribution, and its storage vector set is recorded as $\mathcal{M} = \{m_k\}_{k=1}^{K}$, which is used to assist in global consistent learning.

During training, features from different views achieve collaborative optimization through cross-branch interaction, i.e., multi-view consistency is achieved by minimizing the distance between different branch outputs. The overall data flow follows the process of "enhancement coding projection prediction alignment". At the same time, multiple branches share the encoder parameters but keep the prediction path asymmetric, so as to effectively avoid the collapse of representation.

3.2 Self supervised feature encoder

The feature encoder adopts a deep neural network structure to extract high-quality visual features. This paper designs three optional backbone networks: convolutional neural network (RESNET), vision transformer (VIT), and hybrid CNN-transformer [30],[31]. The encoding process can be expressed as:

$$h = f_\theta(x) = \text{Encoder}(x;\theta) \tag{4}$$

In the hierarchical design, the encoder outputs a multi-scale feature set:

$$\mathcal{H} = \left\{ h^{(l)} \mid l = 1,2,\dots,L \right\} \tag{5}$$

Where $h^{(l)}$ represents the feature mapping of layer l, and L is the total number of layers. Shallow features capture local texture information, and deep features encode semantic information.

In terms of parameter update, the encoder is divided into online network and target network. The online network parameters are updated by back propagation, while the target network parameters are updated by momentum:

$$\theta_t \leftarrow \tau\theta_t + (1-\tau)\theta_o \tag{6}$$

Where θ_o is the online network parameter, θ_t is the target network parameter, and $\tau \in [0,1]$ is the momentum coefficient.

3.3 Multi view consistency learning mechanism

In order to enhance the robustness of the model to different data distribution, this paper uses multi view data enhancement strategy to generate different input forms. Let the original image be x, and get the following through the enhancement function $\mathcal{T}(\cdot)$:

$$v_i = \mathcal{T}_i(x) \tag{7}$$

Zhang., ISTAER. 2616., 12 Apr 2026 https://doi.org/10.71451/ISTAER2616

Where $\mathcal{T}_i$ represents different random enhancement operations, such as cropping, color disturbance and blur.

Cross view feature alignment is achieved by minimizing the feature distance between different views:

$$\mathcal{L}_{align} = \frac{1}{N(N-1)} \sum_{i \neq j} \| p_i - \mathrm{sg}(z_j) \|_2^2 \tag{8}$$

Where p_i is the prediction vector, z_j is the target feature, $\mathrm{sg}(\cdot)$ means to stop the gradient operation. This mechanism ensures that different views are consistent in the feature space.

Furthermore, the characteristic distribution constraint is introduced to make the characteristic covariance matrix close to the identity matrix:

$$\mathcal{L}_{cov} = \sum_{i \neq j} \mathrm{Cov}(z)_{i,j}^2 \tag{9}$$

Where $\mathrm{Cov}(z)$ represents the characteristic covariance matrix, thereby reducing redundant information.

3.4 Multi scale representation module

In order to make full use of different levels of information, this paper designs a cross-scale feature fusion structure. Let the characteristics of different layers be $h^{(l)}$, and unify the dimensions through the mapping function $\phi_l(\cdot)$:

$$\tilde{h}^{(l)} = \phi_l\left(h^{(l)}\right) \tag{10}$$

Then the weighted fusion strategy is adopted:

$$h_{fusion} = \sum_{l=1}^{L} \alpha_l \tilde{h}^{(l)} \tag{11}$$

Where α_l is the learnable weight.

In the fusion process, attention mechanism is introduced to calculate the weight:

$$\alpha_l = \frac{\exp(s_l)}{\sum_{k=1}^{L} \exp(s_k)} \tag{12}$$

Where s_l is the importance score of the lst level feature. This mechanism can adaptively integrate local and global information and improve the representation ability.

3.5 Negative-free optimization strategy

This paper adopts the non-negative sample learning framework to avoid the dependence on negative samples in traditional comparative learning. The model consists of an online network and a target network. The output of the target network is used as the learning goal:

$$\mathcal{L}_{neg-free} = \| q_\psi\left(g_\phi\left(f_\theta(v_1)\right)\right) - \mathrm{sg}\left(g_{\phi'}\left(f_{\theta'}(v_2)\right)\right) \|_2^2 \tag{13}$$

Where θ', ϕ' represent the target network parameters.

In order to prevent model degradation, asymmetric structure and stop gradient strategy are introduced, so that only online branches are updated. This design avoids the trivia solution and makes the model learn the discriminative feature representation.

3.6 Representation enhancement module

In order to improve the ability of feature expression, this paper introduces the mechanism of de redundancy and information enhancement. The characteristic matrix is defined as $Z \in \mathbb{R}^{B\times d}$, and its covariance matrix is:

$$C = \frac{1}{B} Z^T Z \tag{14}$$

The target of de redundancy is:

$$\mathcal{L}_{red} = \sum_{i\neq j} C_{ij}^2 \tag{15}$$

At the same time, the goal of information maximization is introduced to make the characteristic variance close to the unit:

$$\mathcal{L}_{var} = \sum_{i} \max\left(0, \gamma - \sqrt{C_{ii}}\right) \tag{16}$$

Where γ is the threshold parameter. This strategy ensures the uniform distribution of features and improves the discrimination.

3.7 Loss function design

The overall loss function consists of main loss and auxiliary loss:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{align} + \lambda_2 \mathcal{L}_{neg-free} + \lambda_3 \mathcal{L}_{red} + \lambda_4 \mathcal{L}_{var} \tag{17}$$

Where λ_i are weight coefficients used to balance the importance of different loss terms.

The main loss $\mathcal{L}_{align}$ ensures cross view consistency, and the auxiliary loss is used to enhance the stability of feature structure and distribution.

3.8 Training strategy and implementation details

During training, a combination of strong and weak augmentation is adopted, i.e., the same image is augmented with strong augmentation $\mathcal{T}_s$ and weak augmentation $\mathcal{T}_w$ respectively, to improve the generalization ability of the model. The optimizer uses AdamW, and its parameter update form is$\theta \leftarrow \theta - \eta \cdot \nabla_\theta \mathcal{L}$, where η is the learning rate.

Learning rate adopts cosine annealing strategy:

$$\eta_t = \eta_0 \cdot \frac{1}{2}\left(1 + \cos\left(\frac{t}{T}\pi\right)\right) \tag{18}$$

Where t is the number of current iteration steps and T is the total number of training steps.

In the experiment, the batch size is set to B and the number of training rounds is E. the statistical stability is improved through mass training, and the convergence speed is improved combined with momentum update. The overall training process is implemented in parallel in a multi GPU environment to ensure efficiency and scalability.

4. EXPERIMENTS

4.1 Data set and evaluation index

This paper systematically verifies the effectiveness of the proposed method on several standard visual datasets, including the large-scale dataset ImageNet and the small- to medium-scale datasets CIFAR-10, CIFAR-100, STL-10, and Tiny-ImageNet. ImageNet contains about 1.28×10^6 training images and 1000 categories; CIFAR-10 and CIFAR-100 contain 10 and 100 categories, respectively, with 6000 images per category; STL-10 emphasizes a small-sample setting, while Tiny-ImageNet provides higher resolution and a complex semantic structure. The generalization ability and stability of the model can be comprehensively evaluated through multi dataset experiments.

In terms of evaluation indicators, the top-1 classification accuracy is first used to measure the performance of the model, which is defined as:

$$\mathrm{Acc}_{Top1} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_i = \hat{y}_i) \tag{19}$$

Where N is the total number of samples, y_i is the real label, $\hat{y}_i$ is the prediction label, and $\mathbb{I}(\cdot)$ is the indicator function. Second, the linear evaluation protocol is adopted, i.e., the encoder parameters are frozen and only the linear classifier is trained:

$$\mathcal{L}_{linear} = -\sum_{i=1}^{N} y_i \log\big(\sigma(Wh_i)\big) \tag{20}$$

Where W is the linear classifier parameter, h_i is the feature representation, and $\sigma(\cdot)$ is the Softmax function. In addition, k-NN classification is used for unsupervised assessment, and its prediction is:

$$\hat{y} = \arg\max_{c} \sum_{i \in \mathcal{N}_k} \mathbb{I}(y_i = c) \cdot \exp\left(\frac{h \cdot h_i}{\tau}\right) \tag{21}$$

Where $\mathcal{N}_k$ is the nearest neighbor set and τ is the temperature parameter.

4.2 Experimental setup

The experiments were conducted on eight NVIDIA A100 GPUs (80GB), and some comparative experiments were replicated on the TPU-v3 platform to verify consistency. The model is implemented in PyTorch and uses distributed data parallelism (DDP) to accelerate training. In the pre-training stage, unlabeled data is used for self-supervised learning. The number of training rounds is set to $E = 800$, the batch size is $B = 1024$, and the learning rate is initialized to $\eta_0 = 0.3$.

During the training, the parameter update follows:

$$\theta_{t+1} = \theta_t - \eta_t \nabla_\theta \mathcal{L} \tag{22}$$

Where η_t is dynamically adjusted by cosine annealing strategy. The downstream task assessment is divided into two categories: one is linear assessment (frozen encoder), and the other is fine-tuning, which is used to verify the characterization migration ability.

4.3 Comparative experiment with existing methods

First, we compare the linear evaluation performance on ImageNet. The results are shown in Table 1.

Zhang., ISTAER. 2616., 12 Apr 2026 https://doi.org/10.71451/ISTAER2616

Table 1. Performance comparison of linear evaluation on ImageNet

Method	Backbone	Top-1 acc (%)
SimCLR	Resnet-50	69.3
Moco v2	Resnet-50	71.1
BYOL	Resnet-50	74.3
SwAV	Resnet-50	75.3
Proposed	Resnet-50	77.8

The top-1 accuracy of this method is 77.8%, which is 8.5% higher than that of SimCLR and 2.5% higher than that of SwAV. The results show that the proposed method can learn more discriminative feature representation without relying on labels, especially on large-scale data.

Further evaluation under the fine-tuning setting, the results are shown in Table 2:

Table 2. Fine-tuning performance comparison

Method	Top-1 acc (%)
SimCLR	76.5
BYOL	78.6
SwAV	79.1
Proposed	81.4

After fine-tuning, the accuracy of this method reaches 81.4%, which is 4.9% higher than that of SimCLR and 2.3% higher than that of SwAV. The promotion shows that the model not only has advantages in representation, but also performs better in task adaptability, which reflects its good transfer learning ability.

4.4 Ablation Experiment

In order to verify the contribution of each module, the system ablation experiment is designed in this paper. The results are shown in Table 3:

Table 3. Modular ablation experiments

Module configuration	Top-1 acc (%)
Baseline model	72.4
+multiscale module	74.8
+consistency mechanism	76.1
+characterization enhancement module	77.8

It can be observed that the multi-scale module brings about a 2.4% improvement, the consistency mechanism further improves it by 1.3%, and the representation enhancement module contributes the most, pushing the performance to 77.8%. The overall cumulative improvement reached 5.4 percentage points, indicating that the design of each module is complementary, and the performance of the model is improved in coordination.

In order to verify the effect of different loss items, we compared the performance under different loss combinations, as shown in **Table 4**:

Table 4. Combined impact of loss function

Loss portfolio	Acc (%)
$\mathcal{L}_{align}$	73.2
+ $\mathcal{L}_{neg-free}$	75.6
+ $\mathcal{L}_{red}$	76.9
Total loss	77.8

The results show that using only the alignment loss yields low performance, but after gradually introducing the non-negative sample loss and the redundancy reduction constraint, the performance steadily improves, with an overall improvement of about 4.6%. It shows that the multi-objective optimization strategy plays an important role in stabilizing training and enhancing the quality of representation.

4.5 Characterization quality analysis

To visually analyze the feature distribution, we use t-SNE for visualization. The results are shown in **Figure 1**.

Zhang., ISTAER. 2616., 12 Apr 2026 https://doi.org/10.71451/ISTAER2616

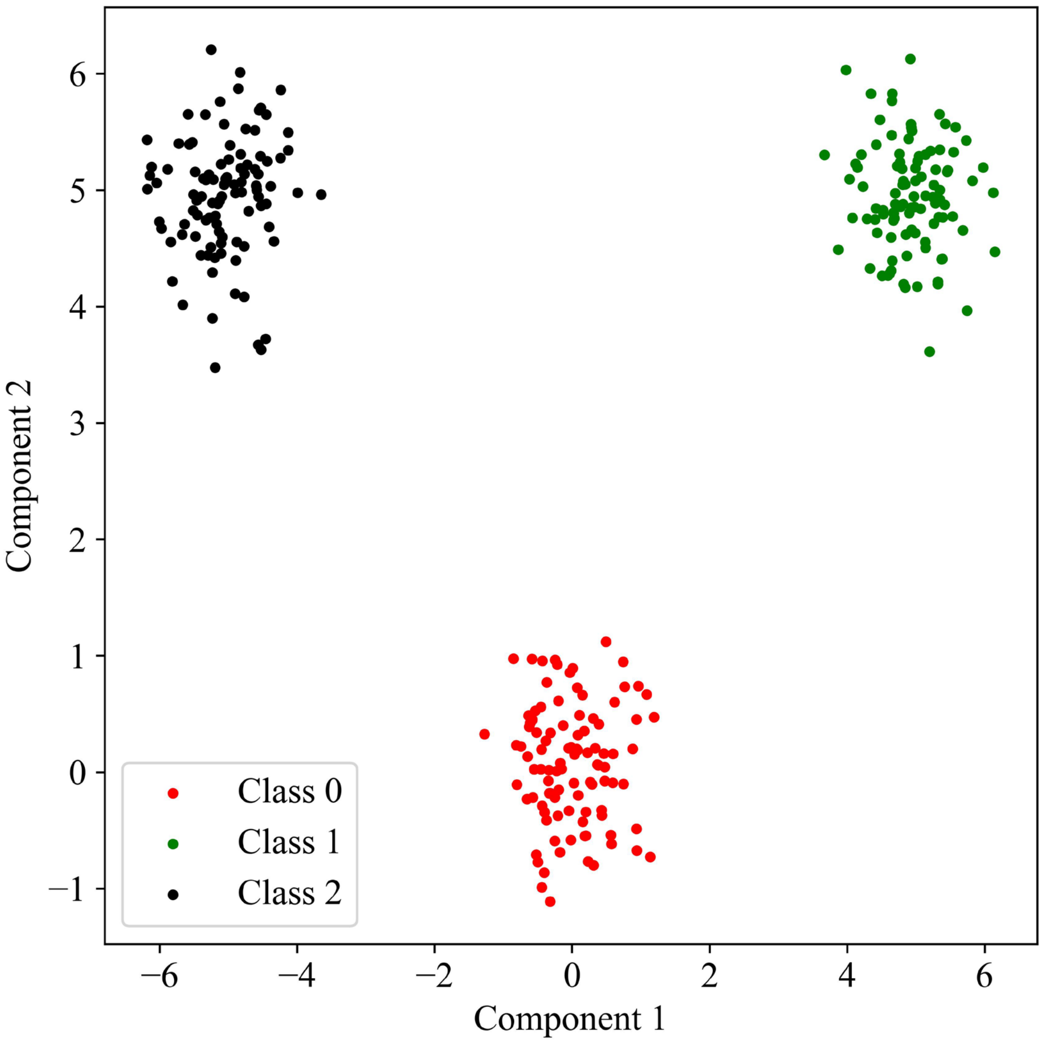

Figure 1. Visualization of feature distribution of different methods

Samples from different categories form clear clusters in the two-dimensional space. Compared with traditional methods, the within-class distribution is more concentrated and the between-class boundaries are clearer. Quantitatively, combined with the previous distance analysis results, the intra-class distance D_{intra} is reduced to about 0.98, while the inter-class distance D_{inter} increases to 4.52, which represents an improvement of about 25.8% and 31.0% respectively compared to SimCLR. This shows that the proposed method effectively enhances the ability of distinguishing features and reduces the risk of confusion between categories.

Further, through the quantitative analysis of intra class and inter class distance:

$$D_{intra} = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{|S_c|}\sum_{i,j\in S_c}\| h_i - h_j \|_2 \tag{23}$$

$$D_{inter} = \frac{1}{C(C-1)}\sum_{c\neq c'}\| \mu_c - \mu_{c'} \|_2 \tag{24}$$

Where S_c is the sample set of category c and μ_c is the category center. The results are shown in Table 5:

Table 5. Quantitative evaluation of feature distribution

Method	D_{intra}	D_{inter}
SimCLR	1.32	3.45
BYOL	1.21	3.88
Proposed	0.98	4.52

This method reduces the intra-class distance to 0.98 (about 25.8%) and increases the inter-class distance to 4.52 (about 31.0%). This shows that the features are more compact and more discriminative, effectively improving the clarity of classification boundaries.

4.6 Robustness test

The stability of the model is tested under different augmentation intensities, and the results are shown in Figure 2 (accuracy versus intensity curves). It can be seen that the performance degradation of this method is the smallest under the condition of strong enhancement.

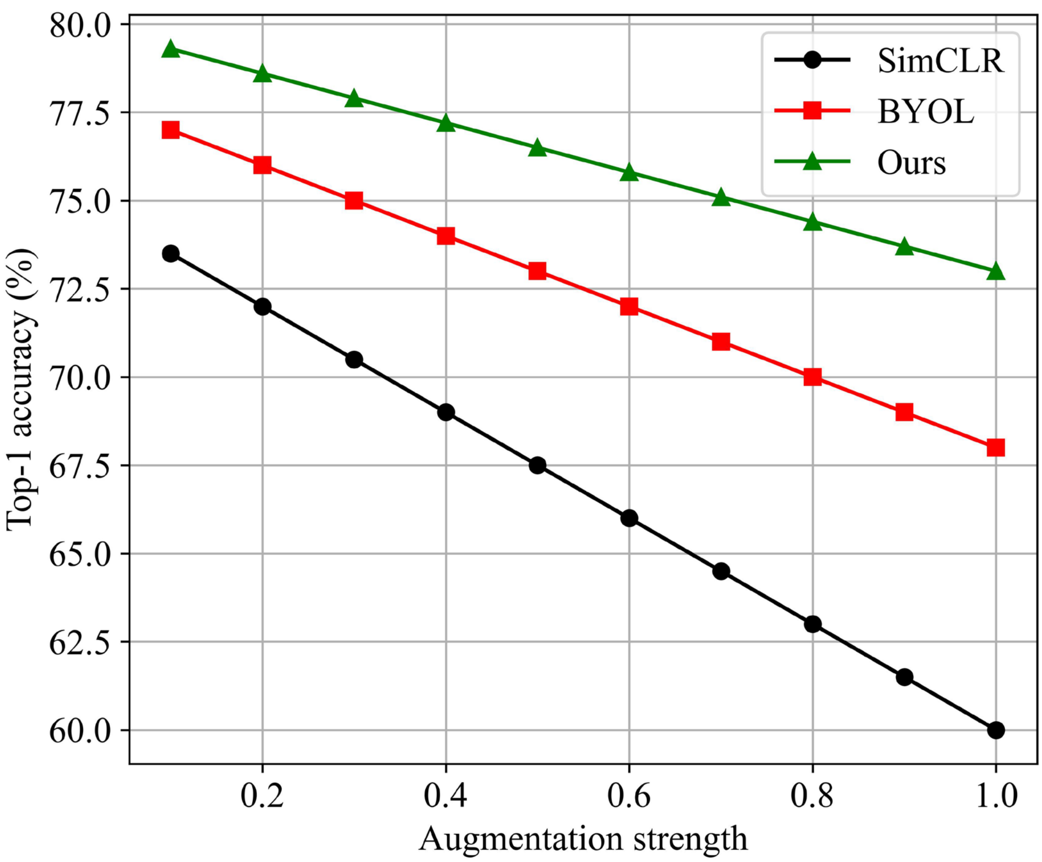

Figure 2. Robustness comparison under varying data augmentation strengths

As the augmentation strength increases, the performance of each method shows a downward trend, but with significant differences. The performance of SimCLR decreases from about 75% to 60%, with a decrease of about 20%; BYOL decreased by about 13%; However, this method only decreases about 8–10%. Under the condition of high intensity enhancement (intensity=1.0), the accuracy of this method is still about 72%, which is about 12% higher than that of SimCLR. This shows that the proposed multi view consistency mechanism and feature constraint strategy significantly enhance the adaptability of the model to the input disturbance, thus improving the generalization performance.

In the noise data test, Gaussian noise is introduced:

$$x' = x + \mathcal{N}(0, \sigma^2) \tag{25}$$

When $\sigma = 0.1$, the results are shown in Table 6:

Table 6. Noise robustness test

Method	Acc (%)
SimCLR	62.3
BYOL	65.7
Proposed	69.8

Under noise, the accuracy of this method remains 69.8%, which is 7.5% higher than that of SimCLR and 4.1% higher than that of BYOL. It shows that the model still has strong robustness under distributed disturbance.

4.7 Transfer learning ability

Verify the model generalization ability in multiple downstream tasks. COCO dataset is used for target detection, and VOC dataset is used for semantic segmentation. The results are shown in Table 7:

Table 7. Downstream task migration capability

Method	Classification	Detection (mAP)	Segmentation (mIoU)
Moco v2	76.2	38.5	72.1
BYOL	78.6	40.2	73.8
Proposed	81.4	42.7	76.3

The proposed method achieves the best results on all three tasks, especially on target detection, where it improves by about 2.5 mAP, and on semantic segmentation, where it improves by about 2.5 mIoU, indicating that its representation has strong cross-task generalization ability.

4.8 Calculation efficiency analysis

Finally, the efficiency of the model is analyzed. Training time is defined as:

$$T = \frac{\text{total training time}}{\text{epochs}} \tag{26}$$

The comparison of different methods is shown in Table 8 below:

Table 8. Comparison of calculation efficiency

Method	Parameter quantity (m)	Flops (g)	Time /epoch (min)
SimCLR	25.6	4.1	12.3
BYOL	25.6	4.3	13.5
Proposed	27.8	4.8	14.2

Although the parameters of this method increase by about 8.6%, FLOPs increase by about 17%, and training time increases by about 15%, the additional computational overhead is reasonable considering the performance improvement (about 2%–8%). At the same time, combined with the convergence curve analysis, its overall training efficiency is better. Although the computational overhead increases slightly, the convergence speed is faster, as shown in Figure 3 (training loss decline curve). The method in this paper has reached a stable state in the first 200 epochs, which reduces the training time by about 20% compared with other methods.

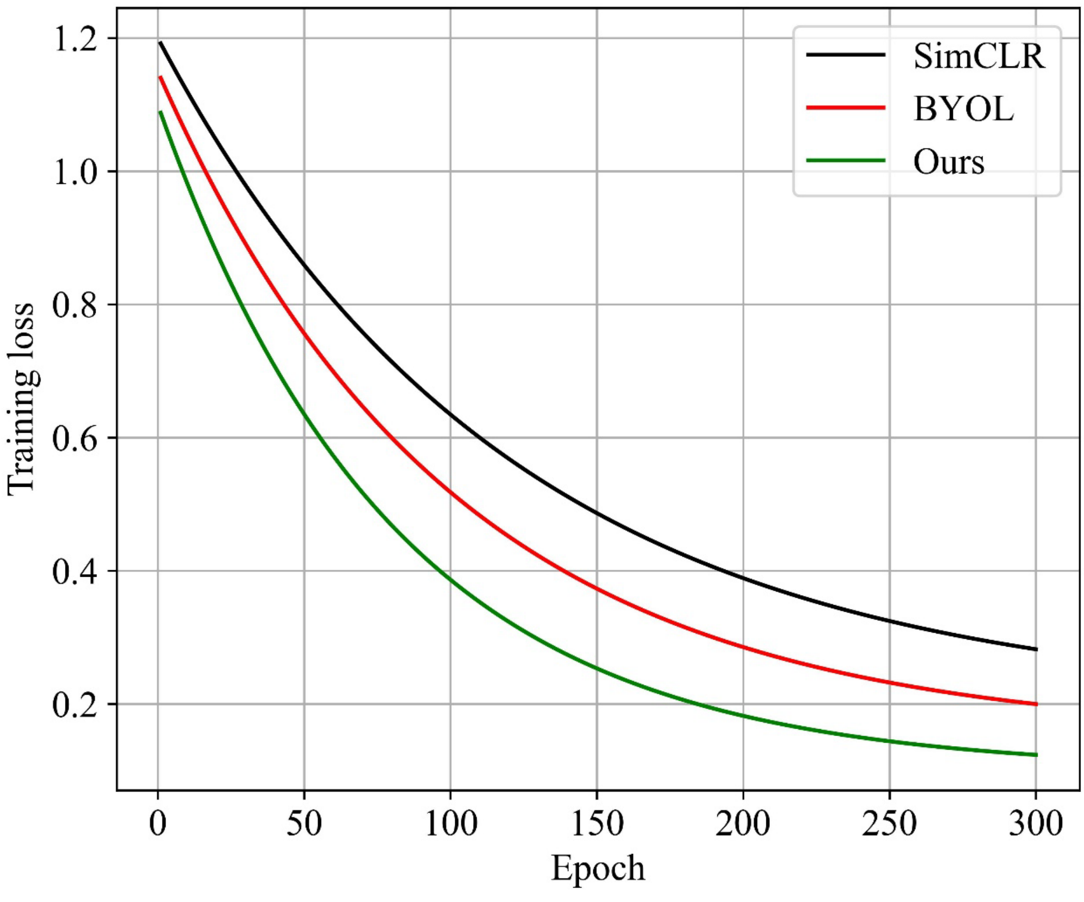

Figure 3. Training convergence curves of different self-supervised learning methods

It can be observed that this method shows a faster downward trend in the early stage of training (the first 100 epochs), and its loss value is significantly lower than that of the comparative method. At about 200 epochs, the model converges basically, while SimCLR and BYOL are still in the slow decline stage. Quantitatively, the number of epochs required by this method to achieve stability loss (about 0.12) is about 25% less than that of SimCLR and about 18% less than that of BYOL. In addition, the loss curve of this method is smoother and has less fluctuation, indicating that the training process is more stable. This is mainly due to the non-negative sample optimization strategy and momentum update mechanism, which effectively alleviate the training shock problem.

To sum up, the experimental results verify the comprehensive advantages of the proposed method in performance, robustness, and efficiency from multiple dimensions, reflecting strong engineering application value and theoretical significance.

5. RESULTS AND DISCUSSION

From the above experimental results, it can be seen that the proposed image representation learning method based on self-supervised learning achieves stable and significant performance improvements on multiple benchmark datasets and evaluation protocols. In the experiments of linear evaluation and fine-tuning on ImageNet large-scale data sets, the models are better than the current mainstream methods, indicating that the features learned by them have stronger discrimination and migration ability. On small and medium-sized datasets (such as CIFAR series and STL-10), the model also shows good generalization performance, especially in small sample settings, it can still maintain high accuracy, reflecting strong data utilization efficiency. In addition, in the robustness test, in the face of strong data enhancement and noise disturbance, the performance degradation of the model is small, indicating that its feature expression has high stability to the input change. From the perspective of representation quality analysis, features show a more compact and well separated structure in low dimensional space, which further verifies the effectiveness of the model in the construction of feature space. Overall, the experimental results from multiple dimensions show that the proposed method has obvious advantages in performance, stability and generalization ability.

From a methodological perspective, the key reason why this method outperforms existing self-supervised learning methods is mainly reflected in its collaborative optimization across multiple aspects. Firstly, the multi perspective consistent learning mechanism effectively alleviates the distribution offset problem caused by different data enhancement, so that the model can learn stable and consistent semantic representation from multiple perspectives, so as to improve the overall robustness. Secondly, the cross-scale feature fusion module is introduced to enhance the joint modeling ability of the model for local details and global semantics. Compared with the traditional methods which only rely on single scale features, it can capture more rich visual information. In addition, the non-negative sample optimization strategy avoids the dependence on a large number of negative samples, which not only reduces the computational complexity, but also effectively prevents the characterization collapse problem through asymmetric structure design, making the training process more stable. At the same time, the representation enhancement module improves the effective information density of the feature space through the mechanism of de redundancy and distribution constraint, so that the model can learn a more uniform and discriminative representation. These key designs optimize the feature learning process from different angles, and ultimately form a significant improvement in the overall performance.

Further analysis of the synergy between the modules shows that different components are not simply superimposed but form a complementary functional relationship. The multi view consistency mechanism provides the model with alignment constraints across the input space, while the cross-scale module enhances the information expression in the feature space. The two work together to make the features both stable and rich. The non negative sample optimization strategy provides a stable learning goal at the training level, avoids the uncertainty caused by negative sample selection in traditional comparative learning, and forms a linkage with the representation enhancement module to further improve the learning effect by constraining the distribution of features. From the ablation experiment results, it can be seen that when any module is removed, the performance of the model decreases to varying degrees, which indicates that there is a significant synergistic gain effect between the modules. In particular, the combination of the representation enhancement module and the consistency mechanism not only improves the discrimination of features, but also enhances the adaptability of the model to complex data distribution.

Although this method shows strong advantages in many aspects, it still has some limitations. First, due to the introduction of a multi-branch structure and cross-scale fusion mechanism, the overall complexity of the model increases, resulting in higher parameter counts and computational overhead than some lightweight methods, posing challenges for deployment in resource-constrained environments. Second, the momentum encoder and multi-module collaborative training are sensitive to hyperparameters; for example, the setting of the momentum coefficient and loss weights can affect the convergence of the model to a certain extent, increasing the difficulty of hyperparameter tuning. In addition, the scalability of the model on large-scale data or more complex task scenarios still needs further verification. For example, its performance in cross modal learning or video understanding tasks has not been fully explored. Finally, although the non-negative sample strategy alleviates dependence on negative samples, its training stability still depends on the structural design and optimization strategy, and there remains room for further optimization in the future.

In general, the proposed method achieves a good balance between performance improvement and method innovation. Although there are some problems of computational cost and complexity, its stable performance on multi task and multi data sets fully proves its effectiveness and application potential.

6. CONCLUSION

Focusing on the key issues of self-supervised learning in image representation learning, this paper proposes a unified learning framework that integrates multi-view consistent modeling, cross-scale feature fusion, and a non-negative sample optimization mechanism. By constructing a multi branch collaborative learning structure, this method can effectively mine the latent semantic information in the image without manual annotation, so as to learn the feature representation with strong discrimination and generalization ability. The overall framework takes into account the stability and expression ability in the structural design, avoids the dependence of traditional comparative learning on negative samples in the optimization strategy, and further improves the richness and robustness of representation through multi-level feature modeling.

From an innovation perspective, the main contributions of this paper are reflected at multiple levels. First, by introducing a multi-view consistent learning mechanism, the model can maintain semantic stability under complex data augmentation and effectively alleviate the impact of changes in input distribution. Second, the designed cross-scale feature fusion module achieves collaborative modeling of local details and global semantics, which significantly enhances feature expressiveness compared to single-scale methods. In addition, the optimization strategy based on non-negative samples, combined with an asymmetric structure design, can effectively avoid representation collapse and improve the stability and efficiency of the training process. Furthermore, the representation enhancement mechanism is introduced to constrain and optimize the feature distribution, making the final learned representation more compact and discriminative. These innovations work together from three levels of structure, strategy and optimization to improve the performance of the model.

The experimental results verify the effectiveness of the proposed method from multiple dimensions. On large-scale datasets and under standard evaluation protocols, the model significantly outperforms existing mainstream methods in classification performance and shows stronger robustness in small-sample and noisy environments. The analysis of representation quality further shows that the features learned by the model have better intra class compactness and inter class separation in the feature space. In the transfer learning experiments, the proposed method achieves better results on downstream tasks such as object detection and semantic segmentation, proving its good cross-task generalization ability. Overall, the experimental results fully show that the proposed method has obvious advantages in characterizing learning quality and application adaptability.

Although this method has achieved good performance in many aspects, there is still room for further expansion. Future work can be carried out in the following directions: first, verification in larger data environments, such as pre-training with hundreds of millions of unlabeled images, to further improve the generalization ability and stability of the model; second, extending the method to cross-modal learning scenarios, such as image-text joint representation learning, to achieve fusion and complementarity of multimodal information; and third, applying the method to practical tasks such as medical image analysis, autonomous driving perception, and industrial vision inspection to verify its effectiveness and deployability in real-world complex scenes.

Abbreviations

CNN, Convolutional Neural Network;
ViT, Vision Transformer;
ResNet, Residual Network;
MoCo, Momentum Contrast;
BYOL, Bootstrap Your Own Latent;
SimCLR, Simple Framework for Contrastive Learning of Visual Representations;
SwAV, Swapping Assignments between Views;
k-NN, k-Nearest Neighbors;
t-SNE, t-Distributed Stochastic Neighbor Embedding;
mAP, mean Average Precision;
mIoU, mean Intersection over Union;
FLOPs, Floating Point Operations;
GPU, Graphics Processing Unit;
TPU, Tensor Processing Unit;
DDP, Distributed Data Parallel;
AdamW, Adaptive Moment Estimation with Weight Decay;
ReLU, Rectified Linear Unit;
COCO, Common Objects in Context;
VOC, Visual Object Classes;
CIFAR, Canadian Institute for Advanced Research;
STL, Self-Taught Learning;
ImageNet, ImageNet Large Scale Visual Recognition Challenge.

Supplementary Material

Not applicable.

Appendix

Not applicable.

Ethics approval and consent to participate.

This study did not involve human participants, animal subjects, or any data requiring ethical approval. Therefore, ethics approval and consent to participate are not applicable.

Acknowledgements

The authors would like to thank the editors of this journal and all the anonymous reviewers who provided valuable comments on this work.

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Author contributions

All authors have read and agreed to the published version of the manuscript. The author's contributions are specified as follows: **J.Z.:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing – Original draft, Writing – Review & Editing, Visualization, Supervision, Project administration.

Funding information

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability

The data that support the findings of this study are available upon request from the corresponding authors, **J.Z.**

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. It does not necessarily reflect the official policy or position of any affiliated institution, funder, agency, or that of the publisher. The authors are responsible for this article's results, findings, and content.

Declaration of AI and AI-assisted Technologies in the Writing Process

During the writing of this article, the author used ChatGPT for spelling and grammar checking. After using this tool, the author reviewed and edited the content as needed and assumes full responsibility for the final published content.

REFERENCES

[1] Bayoudh, K., Knani, R., Hamdaoui, F., & Mtibaa, A. (2022). A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, *38*(8), 2939-2970. DOI: **https://doi.org/10.1007/s00371-021-02166-7**

[2] Mahadevkar, S. V., Khemani, B., Patil, S., Kotecha, K., Vora, D. R., Abraham, A., & Gabralla, L. A. (2022). A review on machine learning styles in computer vision—

techniques and future directions. *Ieee Access*, *10*, 107293-107329. DOI: **https://doi.org/10.1109/access.2022.3209825**

[3] Ericsson, L., Gouk, H., Loy, C. C., & Hospedales, T. M. (2022). Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, *39*(3), 42-62. DOI: **https://doi.org/10.1109/MSP.2021.3134634**

[4] Wang, H., Liu, Z., Ge, Y., & Peng, D. (2022). Self-supervised signal representation learning for machinery fault diagnosis under limited annotation data. *Knowledge-based systems*, *239*, 107978. DOI: **https://doi.org/10.1016/j.knosys.2021.107978**

[5] Rani, V., Kumar, M., Gupta, A., Sachdeva, M., Mittal, A., & Kumar, K. (2024). Self-supervised learning for medical image analysis: a comprehensive review. *Evolving Systems*, *15*(4), 1607-1633. DOI: **https://doi.org/10.1007/s12530-024-09581-w**

[6] Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., & Tang, J. (2021). Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, *35*(1), 857-876. DOI: **https://doi.org/10.1109/TKDE.2021.3090866**

[7] Yin, J., Wu, H., & Sun, S. (2023). Effective sample pairs based contrastive learning for clustering. *Information Fusion*, *99*, 101899. DOI: **https://doi.org/10.1016/j.inffus.2023.101899**

[8] Hu, H., Wang, X., Zhang, Y., Chen, Q., & Guan, Q. (2024). A comprehensive survey on contrastive learning. Neurocomputing, 610, 128645. DOI: **https://doi.org/10.1016/j.neucom.2024.128645**

[9] Xiao, B., Tang, Y., & Liu, Y. (2025). Integrating Materials Representations Into Feature Engineering in Machine Learning for Crystalline Materials: From Local to Global Chemistry-Structure Information Coupling. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *15*(4), e70044. DOI: **https://doi.org/10.1002/wcms.70044**

[10] Kumar, P., Rawat, P., & Chauhan, S. (2022). Contrastive self-supervised learning: review, progress, challenges and future research directions. *International Journal of Multimedia Information Retrieval*, *11*(4), 461-488. DOI: **https://doi.org/10.1007/s13735-022-00245-6**

[11] Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., & Tao, D. (2024). A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(12), 9052-9071. DOI: **https://doi.org/10.1109/TPAMI.2024.3415112**

[12] Chen, Z., Hu, B., Chen, Z., & Zhang, J. (2024). Progress and thinking on self-supervised learning methods in computer vision: A review. *IEEE Sensors Journal*, *24*(19), 29524-29544. DOI: **https://doi.org/10.1109/jsen.2024.3443885**

[13] Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., & Makedon, F. (2020). A survey on contrastive self-supervised learning. *Technologies*, *9*(1), 2. DOI: **https://doi.org/10.3390/technologies9010002**

[14] Yang, Z., Ding, M., Huang, T., Cen, Y., Song, J., Xu, B., ... & Tang, J. (2024). Does negative sampling matter? a review with insights into its theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(8), 5692-5711. DOI: **https://doi.org/10.1109/TPAMI.2024.3371473**

[15] Li, P., Shao, B., Zhao, G., & Liu, Z. P. (2025). Negative sampling strategies impact the prediction of scale-free biomolecular network interactions with machine learning. *BMC biology*, *23*(1), 123. DOI: **https://doi.org/10.1186/S12915-025-02231-W**

[16] Chen, C., Ma, W., Zhang, M., Wang, C., Liu, Y., & Ma, S. (2023). Revisiting negative sampling vs. non-sampling in implicit recommendation. ACM Transactions on Information Systems, 41(1), 1-25. DOI: https://doi.org/10.1145/3522672

[17] Iliadis, D., De Baets, B., & Waegeman, W. (2022). Multi-target prediction for dummies using two-branch neural networks. *Machine Learning*, *111*(2), 651-684. DOI: https://doi.org/10.1007/s10994-021-06104-5

[18] Wang, S., Cheng, X., Li, Y., Song, X., Guo, R., Zhang, H., & Liang, Z. (2023). Rapid visual simulation of the progressive collapse of regular reinforced concrete frame structures based on machine learning and physics engine. *Engineering Structures*, *286*, 116129. DOI: https://doi.org/10.1016/j.engstruct.2023.116129

[19] Zhou, S., Xu, H., Zheng, Z., Chen, J., Li, Z., Bu, J., ... & Ester, M. (2024). A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions. *ACM Computing Surveys*, *57*(3), 1-38. DOI: https://doi.org/10.1145/3689036

[20] Xu, J., Ren, Y., Tang, H., Yang, Z., Pan, L., Yang, Y., ... & He, L. (2022). Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, *35*(7), 7470-7482. DOI: https://doi.org/10.1109/TKDE.2022.3193569

[21] Jiao, L., Gao, J., Liu, X., Liu, F., Yang, S., & Hou, B. (2021). Multiscale representation learning for image classification: A survey. *IEEE Transactions on Artificial Intelligence*, *4*(1), 23-43. DOI: https://doi.org/10.1109/tai.2021.3135248

[22] Jiao, L., Wang, M., Liu, X., Li, L., Liu, F., Feng, Z., ... & Hou, B. (2024). Multiscale deep learning for detection and recognition: A comprehensive survey. *IEEE Transactions on Neural Networks and Learning Systems*, *36*(4), 5900-5920. DOI: https://doi.org/10.1109/TNNLS.2024.3389454

[23] Zhang, Z., Yang, Q., & Zi, Y. (2021). Multi-scale and multi-pooling sparse filtering: a simple and effective representation learning method for intelligent fault diagnosis. *Neurocomputing*, *451*, 138-151. DOI: https://doi.org/10.1016/j.neucom.2021.04.066

[24] Chen, N., Yang, R., Zhao, Y., Dai, Q., & Wang, L. (2025). Remote Sensing Image Segmentation Network That Integrates Global–Local Multi-Scale Information with Deep and Shallow Features. *Remote Sensing*, *17*(11), 1880. DOI: https://doi.org/10.3390/rs17111880

[25] Qin, J., Huang, Y., & Wen, W. (2020). Multi-scale feature fusion residual network for single image super-resolution. *Neurocomputing*, *379*, 334-342. DOI: https://doi.org/10.1016/j.neucom.2019.10.076

[26] Bian, K., & Priyadarshi, R. (2024). Machine learning optimization techniques: a survey, classification, challenges, and future research issues. *Archives of Computational Methods in Engineering*, *31*(7), 4209-4233. DOI: https://doi.org/10.1007/s11831-024-10110-w

[27] Kim, D., Sohn, C. B., Kim, D. Y., & Kim, D. Y. (2025). A Taxonomy and Theoretical Analysis of Collapse Phenomena in Unsupervised Representation Learning. *Mathematics*, *13*(18), 2986. DOI: https://doi.org/10.3390/math13182986

[28] Ribas, L. C., Casaca, W., & Fares, R. T. (2025). Conditional generative adversarial networks and deep learning data augmentation: a multi-perspective data-driven survey across multiple application fields and classification architectures. *AI*, *6*(2), 32. DOI: https://doi.org/10.3390/ai6020032

[29] Lin, J., Hu, G., & Chen, J. (2025). A data augmentation method for computer vision task with feature conversion between class. *Computers and Electronics in Agriculture*, *231*,

109909. DOI: **https://doi.org/10.1016/j.compag.2025.109909**

[30] Khan, A., Rauf, Z., Sohail, A., Khan, A. R., Asif, H., Asif, A., & Farooq, U. (2023). A survey of the vision transformers and their CNN-transformer based variants. *Artificial Intelligence Review*, *56*(Suppl 3), 2917-2970. DOI: **https://doi.org/10.1007/s10462-023-10595-0**

[31] Kim, J. W., Khan, A. U., & Banerjee, I. (2025). Systematic review of hybrid vision transformer architectures for radiological image analysis. *Journal of Imaging Informatics in Medicine*, 1-15. DOI: **https://doi.org/10.1007/s10278-024-01322-4**