A Multi-View Clustering Method for Handling Challenging Samples and Class Imbalance

***Binyu Zhao*[a], *Binxiong Li*[a], *Heyang Gao*[a], *Xi Yu*[a,*], *Quanzhou Luo*[a], *Yujie Liu*[a], *Boyan Zhang*[a], *Haojun Gao*[a], *Yuefei Wang*[b,c]**

[a]*Stirling College, Chengdu University, 2025 Chengdu Rd., Chengdu, Sichuan, China, 610106*

[b]*College of Computer Science, Chengdu University, 2025 Chengdu Rd., Chengdu, Sichuan, China, 610106*

[c]*Key Laboratory of Digital Innovation of Tianfu Culture, Sichuan Provincial Department of Culture and Tourism, Chengdu, China.*

E-mail Address:

202316801313@cdu.edu.cn (Binyu Zhao); libinxiong@stu.cdu.edu.cn (Binxiong Li); gaoheyang@stu.cdu.edu.cn (Heyang Gao); yuxi@cdu.edu.cn (Xi Yu); luoquanzhou@stu.cdu.edu.cn (Quanzhou Luo); liuyujie260@stu.cdu.edu.cn (Yujie Liu); 202416801503@cdu.edu.cn (Boyan Zhang); 202416801402@cdu.edu.cn (Haojun Gao); wangyuefei@cdu.edu.cn (Yuefei Wang);

ABSTRACT

Multi-view clustering (MVC) in real-world settings is challenged by hard samples and class imbalance, which may collapse minority clusters and distort inter-cluster geometry in contrastive representation learning. We propose a unified margin-angle-safety framework derived from a margin-regularized mutual-information lower bound. The framework yields uncertainty-calibrated false-negative screening with soft-negative reweighting, which operates within a provable safe region that preserves angular margins, together with imbalance-aware minority-prototype augmentation with distribution alignment, where a centroid-transport regularizer stabilizes minority clusters while maintaining local-global topology. We further formulate a hierarchical contrastive objective that combines instance-level discrimination with cluster-level contrast and centroid consistency, jointly optimized under the same safety-margin constraints. Theoretically, we establish conditions for the safe use of false negatives, stability of minority centroids, and monotonic growth of inter-cluster angular separation during training. Empirically, the method achieves substantial gains across heterogeneous benchmarks: on RGB-D, ACC and ARI increase by 11.59% and 11.91%; on Cora, ARI improves by 14.95%; and on the Prokaryotic dataset, ACC improves by 10.56%. This framework can serve applications such as cross-modal retrieval, biomedical subgroup discovery, and multi-sensor risk analytics.

The source code for this study is available at: https://github.com/YF-W/MVSIB

Keywords: Multi-view clustering; False and hard negatives; Class imbalance; Contrastive learning

1. Introduction

In recent years, the continuous growth of data scale and heterogeneity has rendered the challenge of learning discriminative representations for structural discovery under unsupervised conditions a central concern. Deep clustering, by jointly optimizing representation learning and clustering objectives, has demonstrated remarkable performance across domains such as images, text, and bioinformatics [1-3]. Leveraging multi-layer neural networks, it effectively models complex nonlinear relationships and outperforms traditional approaches in clustering accuracy, noise robustness, and adaptability to heterogeneous data [4,5].

However, single-view methods, however, are limited to a single feature space, making them vulnerable to noise and missing information [6,7]. By contrast, MVC is capable of aggregating complementary information from different sensing modalities or feature extractors, thereby offering greater potential in structural recovery and noise resistance; when combined with the nonlinear modeling capacity of deep networks, deep multi-view clustering (DMVC) has become a major advancement in this field [8]. Existing MVC approaches can be broadly categorized into shallow representation learning and deep representation learning paradigms [9]. Unlike shallow methods that struggle to effectively address complex nonlinear relationships among views, deep representation learning, through multi-layer neural networks, can automatically learn hierarchical feature representations, thereby surpassing traditional techniques in clustering accuracy, noise robustness, and adaptability to heterogeneous data [10]. Consequently, deep representation learning-based multi-view methods are widely adopted, typically encoding each view with multi-layer neural networks while jointly optimizing objectives in a shared latent space [11]. Studies have shown that models based on multi-view autoencoders achieve superior performance in handwritten digit and face image clustering [12], while fusion strategies incorporating graph neural networks enhance the capacity to process social network and multimodal text data [13].

To further enhance the consistency and discriminative power of cross-view representations, contrastive learning has been extensively incorporated into the DMVC framework. In their work on Contrastive Multiview Coding (CMC), Tian et al. were the first to extend the InfoNCE loss to multi-view settings. By maximizing the consistency of representations of the same instance across different views, CMC effectively harnesses the complementary information among views, significantly improving the quality of image data clustering [14]. The core idea of CMC is to employ contrastive loss to draw positive sample pairs closer and push negative sample pairs farther apart in the latent space, thereby strengthening the discriminative capacity of the representations. Subsequently, Yuan et al. proposed the RPCIC framework, which addresses the common challenge of incomplete views in real-world scenarios by designing a contrastive learning strategy augmented with prototype completion. By integrating a robust prototype completion mechanism into the contrastive loss, their method not only mitigates the negative effects of missing views but also further enhances the clustering accuracy [15]. These works demonstrate that incorporating contrastive learning into the DMVC framework not only allows for a deeper exploration of inter-view consistency but also improves the model’s adaptability to complex and dynamic data environments.

Although instance-discrimination-based contrastive learning can acquire discriminative feature representations in unsupervised settings by distinguishing between positive and negative samples, in the most fundamental paradigm of instance discrimination, samples belonging to the same cluster (sharing the same semantic category) may be mistakenly regarded as negatives, thereby undermining clustering consistency [16]. As illustrated in Figure 1, such false negatives are pulled away from their true clusters, while hard negatives that are close in feature space yet belong to different categories tend to linger near decision boundaries, yielding loose cluster shapes and inter-cluster overlap [17]. To alleviate these issues, previous studies have proposed strategies such as debiasing adjustments to the contrastive loss [18], dynamic gradient correction [19], and hybrid negative sampling [20]. However, most of these approaches assume uniform class distributions and provide insufficient support in scenarios characterized by class imbalance. At present, a systematic contrastive learning framework that explicitly accounts for false negatives and hard negatives remains lacking, and its theoretical soundness has yet to be fully established. Furthermore, existing MVC methods often overlook the problem of uneven sample sizes across clusters. In practical clustering tasks, such imbalance frequently causes models to overfit to large clusters while neglecting small ones, rendering representations of minor categories sparse and ultimately degrading clustering performance [21]. An imbalanced cluster distribution may also shift cluster centroids, further undermining the overall stability and interpretability of clustering [22]. Consequently, there is an urgent need for a general strategy that can simultaneously alleviate sample imbalance while preserving geometric stability within a contrastive learning framework. Particularly in deep multi-view clustering, how to introduce compensation mechanisms to enhance minority cluster representations without inducing cluster centroid drift or disrupting the original inter-cluster margins remains insufficiently characterized in theory and lacks verifiable evidence of stability.

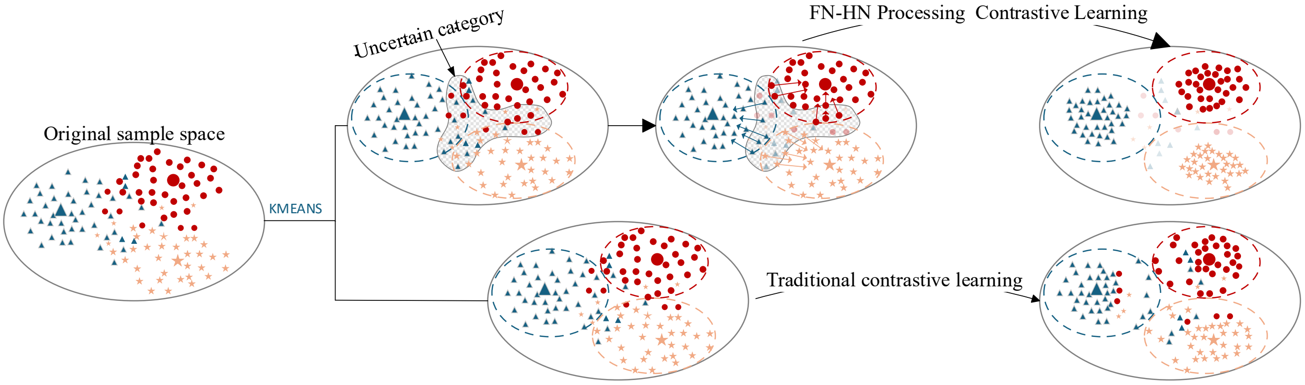


Fig. 1. Solving the Challenging Samples Problem in Contrastive Learning

To address the aforementioned issues, we proposes a unified margin-angle-safety principle to reconcile discriminative power with update safety in contrastive learning. This principle not only unifies instance-level and cluster-level contrastive objectives, but also provides explicit geometric constraints and provable safety guarantees for false-negative handling, minority-cluster compensation, and the repulsion of hard negative samples. Guided by this principle, we design an uncertainty-driven FN/HN discrimination mechanism and a prototype enhancement strategy tailored to minority clusters, and we prove directional safety for updates and stability of minority centroids. Consequently, the traditionally heuristic-driven handling of FN/HN cases and long-tail mitigation is transformed into a learning process endowed with explicit geometric interpretation and provable safety. Based on this, we propose a multi-view contrastive clustering method. In summary, our contributions are as follows:

1) **Margin-angle-safety Principle:** We propose a margin-angle-safety principle that derives both instance- and cluster-level objectives and establishes a provable safe region for using potential false negatives without reducing angular margins.
2) **FN/HN with directional safety:** We propose an uncertainty-calibrated FN/HN discriminator with soft-negative reweighting and prove a directional-safety theorem ensuring updates do not push anchors away from their true clusters.
3) **Hierarchical Contrastive Collaborative Optimization Framework:** We simultaneously minimize both sample-level and cluster-level contrastive objectives: the former suppresses cross-view representation drift, while the latter explicitly enhances "intra-cluster compactness/inter-cluster separation," with adaptive weighting enabling unified convergence from local similarity to global structure.
4) **Minority stability:** We propose minority-prototype augmentation with MMD alignment and boundary constraints, and prove centroid-variance reduction with approximate margin preservation for minority clusters.
5) **Model Advantages and Application Expansion:** Through experimental analysis on five benchmark datasets, we validate the significant advantages of the MVSIB model in the MVC task. The innovation in this study not only advances MVC technology but also provides new perspectives for addressing false negative suppression and class imbalance, with the potential to bring new applications and breakthroughs to fields such as social network analysis and recommendation systems.

Section Arrangement

In Section 2, we systematically review the main and latest advancements in MVC. Section 3 provides a detailed explanation of the proposed model and its theory. In Section 4, we validate the model on multiple public benchmark datasets. Section 5 discusses the limitations and challenges of the proposed method and existing techniques. Finally, Section 6 summarizes the entire work and offers perspectives on potential research directions.

2. Related Work

2.1. Multi-view Clustering

With the proliferation of multi-source heterogeneous data, MVC has gradually emerged as a vital branch of unsupervised learning. Existing approaches can be broadly divided into shallow representation learning and deep representation learning. In shallow methods, researchers typically model each view as a matrix or graph and achieve fusion through algebraic or combinatorial optimization. Representative works include the earliest subspace approaches [23-25], which project multi-view samples into a shared low-dimensional space to obtain consistent representations; graph- and spectrum-based methods [26,27], which construct similarity graphs and employ spectral embedding to capture cross-view local connectivity and cluster structures; multi-kernel clustering [28,29], which leverages complementary kernel functions to adaptively fuse different view-specific metrics in the Reproducing Kernel Hilbert Space (RKHS); and MCLES [30], which jointly solves latent representations and cluster indicator matrices within a unified framework. While these methods have made progress in consistency modeling, they remain dependent on manually designed rules and struggle to address complex nonlinear relationships. To overcome these limitations, deep representation learning methods incorporate the automatic feature extraction capacity of neural networks, making them more suitable for high-dimensional and nonlinear structures. For instance, [31] integrates neural networks with traditional subspace methods, extracting higher-level semantic features through neural networks and applying them to latent space representations; AE2-Nets [32] employ nested autoencoders to integrate heterogeneous sources while balancing consistency and complementarity; furthermore, end-to-end frameworks such as SiMVC, CoMVC [33], CMSC-DCCA [34], and MAGCN [35] introduce self-supervised losses, demonstrating superior performance over traditional methods on large-scale datasets. Overall, whether shallow or deep, existing MVC methods have primarily focused on modeling cross-view consistency and complementarity, yet they have insufficiently exploited discriminative feature learning under unsupervised conditions. In recent years, with the rise of self-supervised contrastive learning, researchers have begun to integrate contrastive paradigms into multi-view clustering, alleviating representation drift and enhancing robustness, thereby bringing about new breakthroughs in this domain.

2.2 Contrastive Learning

In recent years, contrastive learning has gradually been introduced into MVC to enhance cross-view consistency and discriminability. Classic works such as CIRCLE [36] and DMJC [37] have demonstrated the effectiveness of cross-view alignment by constructing positive and negative sample pairs within and across views. However, in unsupervised settings, issues like false negatives and hard negative samples remain prominent. False negatives refer to the misclassification of similar samples as negative. Zhang et al. [38] alleviated this issue by combining node attributes with graph structural similarity detection, while Chien et al. [39] employed an out-of-distribution (OOD) mechanism based on embedding distances for filtering. In this direction, our previous work, MPCCL [40], improved structural preservation and feature diversity through multi-scale graph coarsening and a one-to-many contrastive mechanism. However, its positive and negative sample selection still relied on static partitioning, which failed to effectively avoid the interference of false negatives and overlooked the challenges posed by class imbalance. The hard negative sample issue, typically characterized by close embedding distances between heterogeneous samples, complicates the contrastive loss optimization. To address this, Robinson et al. [41] proposed a conditional probability-based reweighting approach to mitigate the problem, while Jiang et al. [42] incorporated it explicitly into supervised contrastive loss to enhance discriminability. Additionally, in highly imbalanced scenarios such as medical imaging and anomaly detection, Gao et al. [43] significantly improved the representation ability of minority class samples by refining contrastive loss, view resampling, and data augmentation strategies.

2.3 Limitations

Despite the progress achieved by existing MVC methods in information fusion, representation learning, and clustering performance, several limitations remain. Traditional shallow representation learning approaches (such as subspace clustering and graph clustering) rely on linear assumptions and predefined similarity structures, making them ill-suited for capturing complex nonlinear relationships and higher-order dependencies, and incapable of jointly optimizing clustering and representation learning. Deep representation learning methods based on neural networks, while endowed with powerful expressive capacity, often adopt rigid alignment or unified fusion strategies that neglect inter-view heterogeneity, thereby constraining the model's ability to dynamically assign priorities according to view quality. Moreover, such models are structurally intricate, highly sensitive to hyperparameters, initialization, and data quality, and tend to exhibit weakened discriminative ability under class imbalance. In recent years, contrastive learning has been introduced into MVC, enhancing the capacity for learning both consistency and discrimination; however, challenges persist in the selection of positive and negative samples, particularly the misclassification of false negatives and the disruptive influence of hard negatives during optimization. Although certain studies have attempted to mitigate these issues, a unified framework is still lacking, making it difficult to fully exploit multi-source information for optimal representation and discrimination. Consequently, improving sample discrimination accuracy, alleviating class imbalance, and addressing false negative interference remain promising directions for future research.

3. Methodology

In multi-view clustering tasks, our objective is to learn a function $\varphi: \mathcal{X} \to \mathcal{Y}, \mathcal{Y} = [C] = \{1,2,\ldots,C\}$, and $\mathcal{X} = \{(x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(V)})\}_{i=1}^{N}$ N represents the multi-view input space consisting of N samples, with each sample having corresponding feature representations $x_i^{(v)}$ across V views. The goal is to partition these samples into C clusters and output hard cluster labels such that samples within the same cluster are more compact in the consensus space, while samples from different clusters are more distinct. To achieve this, the MVSIB model decomposes the clustering function φ into a deep autoencoder encoder $f^{(v)}: \mathcal{X}^{(v)} \to \mathcal{H}^{(v)} \subseteq \mathbb{R}^h$ and a shared consensus space projection $\Pi: \mathcal{H}^{(v)} \to \mathcal{Z} \subseteq \mathbb{R}^D$, combined with cluster-level contrastive loss and regularization. By training the encoder $f^{(v)}$, the model not only integrates complementary information from multiple views but also generates more discriminative clustering representations while addressing issues such as false negatives, hard negatives, and class imbalance. Table 1 summarizes the main variables and their definitions involved in the MVSIB framework.

3.1 Notation

Table 1. Notation and Definitions

Notation	Explanation
$z_i^{(v)}, z_i^{\text{cons}} \in \mathbb{R}^D$	The latent representation of the i-th sample in the v-th view and its corresponding latent representation in the consensus space.
C	The number of clusters.
$u_c^{(v)}, u_c^{cons} \in \mathbb{R}^D$	The cluster center of the c-th cluster in the v-th view and the cluster center of the c-th cluster in the consensus space.
$n_c \in \mathbb{N}^+$	The number of samples in the c-th cluster.
V	The number of views.
$\sigma_c^{(v)} > 0$	The Gaussian kernel width of the c-th cluster in the v-th view.
$m_{ic}^{(v)}$	The soft assignment probability of sample i to cluster c in the v-th view.
$d_{ic}^{(v)}$	The Euclidean distance between sample i and cluster center c in the v-th view.
$\widetilde{m}_{ic}^{(v)}$	Unnormalized membership degree
$H_i^{(v)}$	The Shannon entropy of the i-th sample in view v.
$h_i^{(v)}$ 、 $\delta_i^{(v)}$ 、 $u_i^{(v)}$ 、 u_i	Normalized entropy, normalized difference, view-level uncertainty, and final fusion uncertainty.

Continued Table 1

U_t、C_t	The current batch of uncertain set; certain set.
FN, HN	False negative set and hard negative set.
p_e	The proportion of uncertain samples in the e-th round.
$\psi_{c,p}$	The p-th pseudo-sample generated from the c-th cluster.
M	The set of minority sample clusters.
Σ_c, Σ'_c	The covariance matrix of the c-th cluster before and after adjustment.
y_{ij}^{pse}、y_i、$\hat{y}_i$	Cluster labels (pseudo-labels), discriminative labels (FN/HN discrimination), and weight matrix (used in contrastive learning).
r	Sample index

3.2 Overall framework

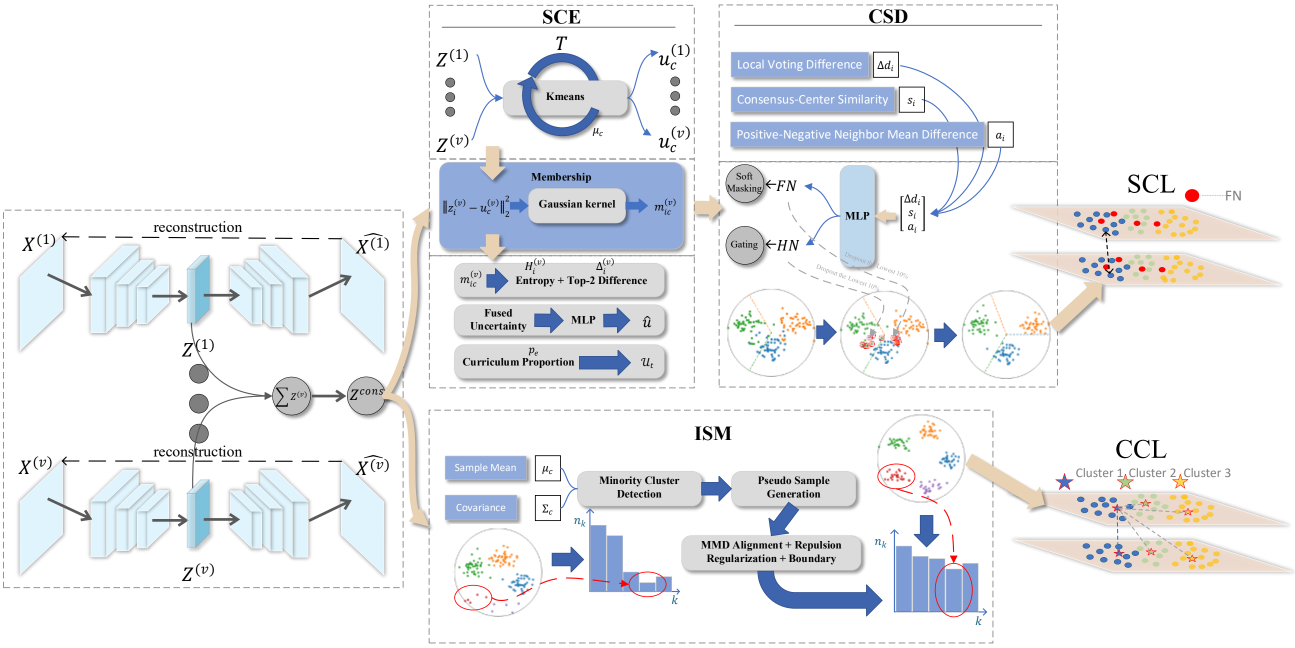


Fig. 2. Overall Framework Architecture

The overall framework of MVSIB is illustrated in Figure 2. Guided by the margin-angle-safety principle, MVSIB is organized around a single objective that enlarges inter-cluster angular margins while enforcing update safety. Concretely, view-specific autoencoders first produce latent embeddings for each view; these are then optimized by three cooperating modules under shared safety-margin constraints. **(i) Instance-level safety (SCE/CSD)**: an uncertainty-aware FN/HN discriminator estimates sample reliability and applies differentiated treatment to uncertain pairs through soft reweighting and push-pull constraints. **(ii) Cluster-level stability (ISM)**: minority-prototype augmentation with distribution alignment and a boundary constraint reduces centroid variance and preserves margins, stabilizing long-tailed clusters. **(iii) Hierarchical contrast (SCL/CCL)**: instance-level discrimination is coupled with cluster-level contrast and centroid consistency, and the two levels are jointly optimized under the same safety-margin constraints. Training proceeds in stages: autoencoder pretraining for view fidelity, a short safety-calibration warm-up, and joint optimization with a two-time-scale update (centers slower than embeddings), which yields monotonic growth of inter-cluster angular separation while fusing complementary multi-view information. The result is a common representation space with tight intra-cluster structure and clear inter-cluster boundaries, robust to false/hard negatives and class imbalance.

3.3 Network Architecture and Consensus Graph Construction

To achieve effective information integration across multiple views, we construct an autoencoder network for each view v. Specifically, the input sample $x^{(v)} \in \mathbb{R}^{d_v}$ of the v -th view is mapped to a latent representation through the encoder:

$$z^{(v)} = f_{\theta^{(v)}}\left(x^{(v)}\right) \tag{1}$$

Here, $f_{\theta^{(v)}}(\cdot)$ denotes the encoder of the v -th view, with $\theta^{(v)}$ representing its parameters. To prevent information loss, each encoder is paired with a symmetric decoder $g_{\phi^{(v)}}(\cdot)$, which minimizes the reconstruction loss to ensure that the latent representation retains sufficient original information.

$$L_{\mathrm{re}} = \sum_{v=1}^{V} \parallel x^{(v)} - g_{\phi^{(v)}}\left(z^{(v)}\right) \parallel^2 \tag{2}$$

Building upon this, we obtain the consensus representation of the sample through cross-view weighting:

$$z^{cons} = \sum_{v=1}^{V} \alpha_v z^{(v)} \tag{3}$$

Here, α_v represents the weight of the v -th view. The weights are obtained by normalizing the learnable parameters w_v through the softmax function.

$$\alpha_v = \frac{\exp(w_v)}{\sum_{j=1}^{V} \exp\left(w_j\right)}, \sum_{v=1}^{V} \alpha_v = 1, \alpha_v > 0 \tag{4}$$

This approach ensures that the contribution of different views to the consensus representation can be adaptively adjusted. The consensus representation not only serves as the foundation for subsequent pseudo-label generation and clustering but also acts as an anchor point for cross-view alignment, constraining the consistency between each view's representation and the consensus space during the contrastive learning phase. To capture the local structure between samples, we further construct a consensus graph based on the consensus representation. Specifically, for small-scale datasets, we compute the k-nearest neighbor graph over the entire dataset using z^{cons}, yielding a global adjacency matrix $W \in \mathbb{R}^{N \times N}$. For large-scale datasets, due to memory constraints, the adjacency matrix is dynamically constructed within each batch.

To ensure that the cluster centers accurately reflect the dynamic distribution of the feature space throughout the training process, we adopt a K-Means-based dynamic update strategy. In the initialization phase, we run K-Means on the latent representations $\left\{z_i^{(v)}\right\}$ for each view v to obtain the cluster centers$\{u_c^{(v)}\}_{c=1}^{C}$for the different views. Subsequently, every t steps, K-Means is rerun on the corresponding view for periodic correction. The update rule is as follows: $u_c^{(v)} \leftarrow \frac{1}{|\Omega_c|} \sum_{i \in \Omega_c} z_i^{(v)}, c = 1, \ldots, C,$where $\Omega_{\mathrm{c}} = \{\, i \mid y_i = c \,\}$ denotes the set of samples assigned to the c-th cluster. Additionally, the consensus cluster center u_c^{cons} is computed, which incorporates information from all views and more accurately reflects the overall cluster structure of the model. The update rule is: $u_c^{cons} =$

$\frac{1}{n_c}\sum_{i:l_i=c} z_i^{cons}, c = 1, \ldots, C$. This strategy ensures that the cluster centers can adaptively aggregate the most recent feature representations and continuously correct the estimated bias of the latent structure during training. In this way, the cluster centers serve both as reference points for positive and negative samples in contrastive learning and enhance the model's ability to delineate semantic boundaries under dynamic feature distributions.

3.4 Sample Certainty Estimation (SCE)

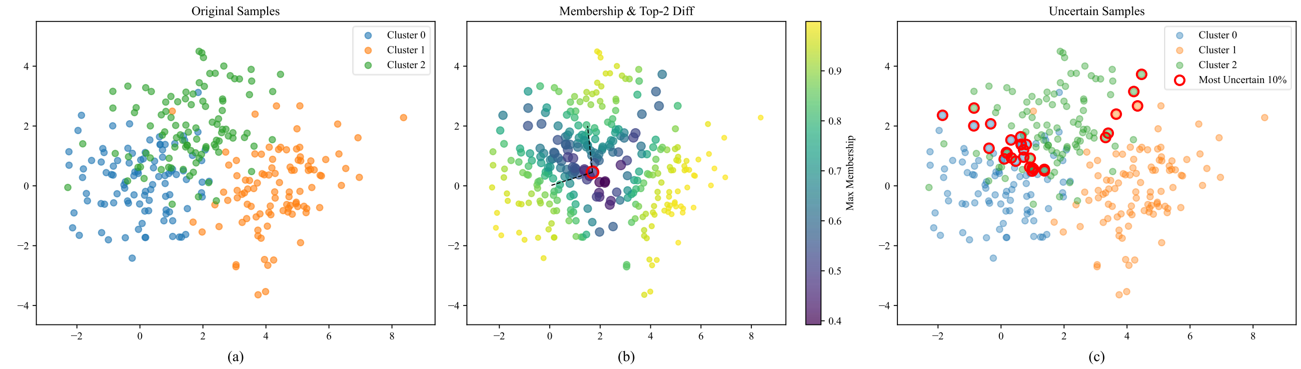


Fig. 3. Visualization of Sample Certainty Estimation (SCE)

The Sample Certainty Estimation module (SCE) quantitatively evaluates the “credibility” of each training sample with respect to the current cluster structure, dynamically categorizing samples into “certain” and “uncertain” groups. The model then leverages this categorization to dynamically reweight the loss and apply corresponding strategies for handling FN and HN. SCE consists of three key steps: view-level membership inference, dual-criterion uncertainty measurement, and curriculum-based sample partitioning. To effectively visualize this process, Figure 3 illustrates the spatial distribution of “certain” and “uncertain” samples inferred by the model within a simulated dataset. In the figure, color and size are used to represent a sample’s cluster membership and its degree of uncertainty, respectively. By jointly measuring the Top-2 probability difference and entropy, the model is able to clearly identify those samples exhibiting the highest uncertainty during clustering. These uncertain samples typically lie at cluster boundaries or within overlapping regions, and are dynamically classified into the uncertain group, thereby providing a foundation for subsequent sample-handling strategies.

3.4.1 View-specific Membership Inference

Let the latent representation of the i-th sample in the v-th view be denoted as $z_i^{(v)}$, with C predefined cluster centers $\{u_c^{(v)}\}_{c=1}^{C}$ and corresponding bandwidths $\{\sigma_c^{(v)}\}_{c=1}^{C}$. The sample-to-cluster distance is first defined as:

$$d_{ic}^{(v)} = \left\| z_i^{(v)} - u_c^{(v)} \right\|_2^2 \tag{5}$$

This distance is then mapped to an unnormalized membership score via a Gaussian kernel:

$$\tilde{m}_{ic}^{(v)} = \exp\left[-\frac{d_{ic}^{(v)}}{2(\sigma_c^{(v)})^2}\right] \tag{6}$$

To obtain a probabilistic soft partitioning, the scores are normalized across all clusters:

$$m_{ic}^{(v)} = \frac{\widetilde{m}_{ic}^{(v)}}{\sum_{j=1}^{C} \widetilde{m}_{ij}^{(v)}}, \sum_{c=1}^{C} m_{ic}^{(v)} = 1 \tag{7}$$

Both $d_{ic}^{(v)}$ and $m_{ic}^{(v)}$ produce differentiable cluster membership distributions for each sample during training, serving as direct inputs for constructing two types of uncertainty measures: entropy-based and separability-based.

3.4.2 Dual-Criterion Uncertainty Measurement

We employ view entropy to capture the dispersion of membership distributions, and boundary risk to evaluate the fuzziness of decision boundaries. By balancing these two measures, we derive the uncertainty of each sample and classify them into either the certain class or the uncertain class. To quantify the degree of dispersion in membership distributions across clusters, we adopt Shannon entropy:

$$H_i^{(v)} = -\sum_{c=1}^{C} m_{ic}^{(v)} \ln\left(m_{ic}^{(v)} + \varepsilon\right), h_i^{(v)} = \frac{H_i^{(v)} - \underset{j}{min} H_j^{(v)}}{\underset{j}{max} H_j^{(v)} - \underset{j}{min} H_j^{(v)} + \varepsilon} \in [0,1], \tag{8}$$

where $\varepsilon > 0$ is a small smoothing constant to prevent numerical instability caused by a zero or excessively small denominator. The min/max operations are performed within the current batch. If $m_{ic}^{(v)}$ is highly concentrated, the entropy tends toward 0, indicating a *strong signal*; if the cluster probabilities are nearly uniform, the entropy approaches $\ln C$, signifying a weak signal. The batch-wise range normalization ensures that $h_i^{(v)}$ maps entropy values to $[0,1]$, thereby stabilizing subsequent weighted fusion across training batches, without being distorted by changes in sample size or cluster number.

We further define the separation degree $\Delta_i^{(v)}$. Let $m_{i(1)}^{(v)} \geq m_{i(2)}^{(v)}$ denote the largest and second-largest components of the membership vector. Then,

$$\Delta_i^{(v)} = m_{i(1)}^{(v)} - m_{i(2)}^{(v)} \tag{9}$$

$$\delta_i^{(v)} = \frac{\Delta_i^{(v)} - \underset{j}{min} \Delta_j^{(v)}}{\underset{j}{max} \Delta_j^{(v)} - \underset{j}{min} \Delta_j^{(v)} + \varepsilon} \in [0,1], \tag{10}$$

When $\Delta_i^{(v)}$ is small, the sample lies close to the boundary between two clusters, raising its uncertainty. To standardize the scale, we use $1 - \delta_i^{(v)}$ as the boundary risk, which also falls within $[0,1]$. Introducing a tunable balance factor $\alpha \in [0,1]$, we obtain:

$$u_i^{(v)} = \alpha h_i^{(v)} + (1 - \alpha_{sce}) \left[1 - \delta_i^{(v)}\right] \tag{11}$$

A larger α_{sce} emphasizes the fuzziness caused by cluster overlap, while a smaller α_{sce} enhances sensitivity to boundary noise.

Since strong ambiguity in any view can sufficiently reflect overall risk, we adopt the maximization criterion:

$$u_i = \underset{v=1,\dots,V}{max} u_i^{(v)} \tag{12}$$

thus yielding the fused uncertainty $u_i \in [0,1]$.

To prevent discarding potentially valuable information too early, we adopt a **linear annealing strategy** for setting the target uncertainty ratio in each training epoch:

$$p_e = p_0 \left(1 - \frac{e-1}{E-1}\right), e = 1, \dots, E \tag{13}$$

$$k_e = max\{1, \lfloor Np_e \rfloor\}, \tag{14}$$

where p_0 is the initial ratio and E denotes the total number of epochs. The predicted uncertainties $\hat{u}_i$ within each batch are ranked in descending order, and the top k_e samples form the $\mathcal{U}_t$, with the remainder classified as the C_t. As e increases, p_e decreases gradually, guiding the model from broad exploration toward refined focus, thereby realizing a course-learning process with progressively reduced difficulty.

3.5 FN Detection and Adaptive Handling of HN Samples (Challenging Sample Disambiguation, CSD)

We subdivide all high-uncertainty samples into FN and HN. FN refers to samples that truly belong to the same cluster but are mistakenly regarded as out-of-cluster due to noise or labeling errors; HN, on the other hand, denotes points originating from different clusters yet lying in close proximity to the target cluster within the latent feature space, thereby prone to confusion. Excessive penalization of FN undermines intra-cluster compactness, whereas neglecting HN diminishes inter-cluster separability. To address these two types of samples, we apply soft masking to FN within the contrastive loss, while intensifying the repulsion of HN, thus maintaining intra-cluster convergence while reinforcing inter-cluster discrimination. To provide an intuitive illustration of this process, Figure 4 visualizes the treatment of FN and HN. Through this refinement, the model achieves a more precise differentiation of high-uncertainty samples, further enhancing cluster boundaries and structures. Consequently, the adverse impact of FN and HN on intra-cluster compactness is effectively mitigated, while inter-cluster separability is improved, thereby strengthening both the clustering performance and stability of the model.

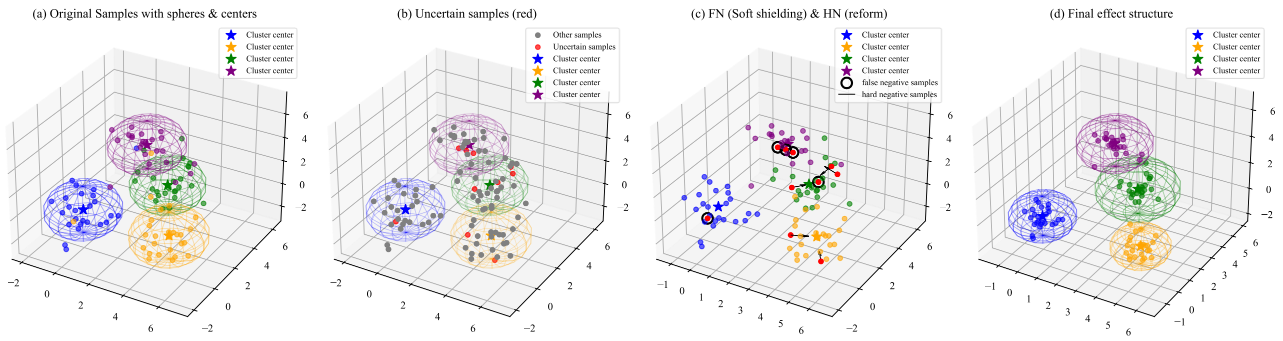


Fig. 4. Visualization of Challenging Sample Disambiguation (CSD)

3.5.1 Constructing Features for FN and HN Sample Discrimination

To accurately distinguish between FN and HN samples from high uncertainty samples, we have designed three complementary indicators: local voting difference Δd_i, global similarity s_i, and cross-view distance difference a_i. These three indicators complement each other across the local, global, and cross-view dimensions, providing a comprehensive quantification of the sample's confusion risk and offering a reliable basis for the subsequent FN/HN segmentation.

(1) Δd_i measures the local "voting difference." In each view v, let $p_i^{(v)}$ and $n_i^{(v)}$ represent the counts of neighbors

in the k-nearest neighbors that fall into the same cluster and different clusters, respectively. Then,

$$\Delta d_i = \frac{1}{V}\sum_{v=1}^{V} \frac{p_i^{(v)} - n_i^{(v)}}{p_i^{(v)} + n_i^{(v)} + \varepsilon} \tag{15}$$

If $\Delta d_i < 0$, the sample is more surrounded by samples from different clusters, indicating a tendency towards HN; conversely, if $\Delta d_i > 0$, the sample is more likely to be a FN.

(2) s_i is the cosine similarity between the sample's consensus representation z_i^{cons} and its cluster center u_c^{cons}.

$$s_i = \cos(z_i^{\mathrm{cons}}, u_c^{cons}) \tag{16}$$

If s_i is relatively high but the sample is still classified as uncertain, it usually suggests a label error, indicating a tendency towards FN.

(3) a_i is the cross-view neighborhood discrimination coefficient. For view v, let $d_i^{\mathrm{pos},(v)}$ and $d_i^{\mathrm{neg},(v)}$ represent the mean distances to the "same-cluster" and "different-cluster" neighbors, respectively. Then,

$$a_i = \frac{1}{V}\sum_{v=1}^{V} \frac{d_i^{\mathrm{neg},(v)} - d_i^{\mathrm{pos},(v)}}{d_i^{\mathrm{neg},(v)} + d_i^{\mathrm{pos},(v)} + \varepsilon} \tag{17}$$

When $a_i > 0$ and the magnitude is large, the sample is closer to the same-cluster center, increasing the probability of it being a FN. Conversely, if $a_i < 0$, the sample tends towards HN.

Based on the above three judgment conditions, we construct a three-dimensional discrimination vector for each uncertain sample $i \in \mathcal{U}_t$.

$$\mathbf{f}_i = [\Delta d_i, \quad s_i, \quad a_i]^\top \tag{18}$$

The discrimination features $\mathbf{f}_i$ are input into a two-layer perceptron (MLP), resulting in a two-dimensional scoring vector.

$$t_i = \begin{bmatrix} z_i^{\mathrm{HN}} \\ z_i^{\mathrm{FN}} \end{bmatrix} = \mathrm{MLP}(f_i) \tag{19}$$

Applying Softmax to z_i yields the posterior probabilities of the sample being classified as HN or FN.

$$p_i^{\mathrm{HN}} = \frac{\exp\left(t_i^{\mathrm{HN}}\right)}{\exp\left(t_i^{\mathrm{HN}}\right) + \exp\left(t_i^{\mathrm{FN}}\right)} \tag{20}$$

$$p_i^{\mathrm{FN}} = \frac{\exp\left(t_i^{\mathrm{FN}}\right)}{\exp\left(t_i^{\mathrm{HN}}\right) + \exp\left(t_i^{\mathrm{FN}}\right)} \tag{21}$$

The final FN/HN label $\hat{y}_i$ for sample i is determined based on the maximum posterior principle:

$$\hat{y}_i = \arg \underset{k\in\{\mathrm{HN,FN}\}}{max} p_i^k = \begin{cases} \mathrm{FN}, \ p_i^{\mathrm{FN}} > p_i^{\mathrm{HN}}, \\ \mathrm{HN, \ otherwise.} \end{cases} \tag{22}$$

Here, $\hat{y}_i \in \{FN, HN\}$ serves as the routing label, which determines whether sample i is routed to the soft-masking branch (FN) or the hard-repulsion branch (HN) for subsequent optimization.

Define the discrimination confidence level.

$$c_i = \max\left\{p_i^{\mathrm{HN}}, p_i^{\mathrm{FN}}\right\} \tag{23}$$

To mitigate misclassifications caused by noisy or model non-convergence, during the early stages of training, the lowest 10% of samples within each batch, ordered by c_i, are excluded and labeled as "undecided." These samples do not contribute to the subsequent differential loss calculations, thereby enhancing the overall stability of the training process.

3.5.2 Design of Differential Loss

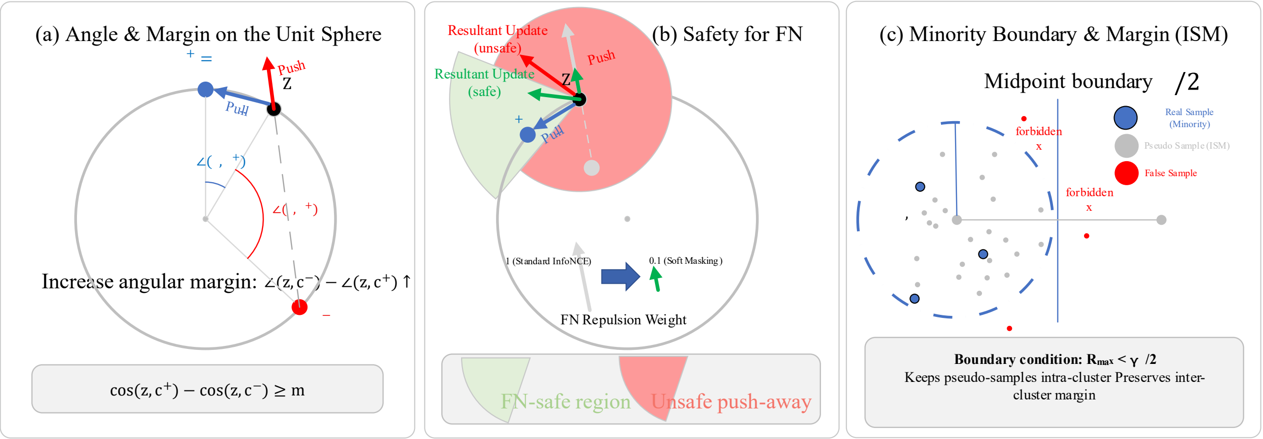
After completing the segmentation of FN and HN, we apply different strategies to the two categories of samples in the contrastive loss. First, the label weight matrix is calculated in the feature consistency term y_{ij}^{pse}.

$$y_{ij}^{\mathrm{pse}}=\begin{cases}1, & c_i=c_j\wedge i\notin \mathrm{FN}\wedge j\notin \mathrm{FN},\\ 0.1, & c_i=c_j\wedge (i\in \mathrm{FN}\vee j\in \mathrm{FN}),\\ 0, & \text{otherwise}.\end{cases}\tag{24}$$

Here, $c_i\in\{1,\dots,C\}$ represents the cluster label of the sample.

3.5.3 Margin-Angle-Safety Framework

Figure 5 provides a geometric overview of our safety-margin principle on the unit hypersphere and previews how it supports the subsequent theoretical guarantees. In Figure 5(a), since all representations are ℓ_2-normalized, cosine similarity corresponds to angular distance on the unit hypersphere, and our margin-based push-pull objective increases inter-cluster separation by enforcing a cosine margin between the assigned centroid c^+ and the nearest non-matching centroid c^-. Figure 5(b) explains directional safety: a false negative u_{FN} mistakenly treated as a negative would generate a push-away component along the divergent direction from the true centroid, whereas our mechanism controls the FN softmax mass θ_i via $w_{FN}=g\pi_{FN}$so that when $\theta_i\le\Theta_i$ the update projection onto the divergent direction is non-positive. Figure 5(c) illustrates minority safety: pseudo-samples are constrained within $R_{\max}$ and kept beyond $R_{\min}$, and the condition $R_{\max}<\gamma/2$ preserves inter-cluster margins while reducing centroid variance.



(a) Angle & Margin on the Unit Sphere
+ =
Push
Z
Pull
∠(, +)
∠(, +)
Increase angular margin: ∠(z, c⁻) − ∠(z, c⁺) ↑
−
cos(z, c⁺) − cos(z, c⁻) ≥ m
Resultant Update (unsafe)
Resultant Update (safe)
(b) Safety for FN
Push
Z
+
Pull
1 (Standard InfoNCE)
0.1 (Soft Masking)
FN Repulsion Weight
FN-safe region
Unsafe push-away
(c) Minority Boundary & Margin (ISM)
Midpoint boundary /2
forbidden x
forbidden x
Real Sample (Minority)
Pseudo Sample (ISM)
False Sample
Boundary condition: Rmax < γ /2
Keeps pseudo-samples intra-cluster Preserves inter-cluster margin

Fig. 5. Margin-Angle-Safety Framework

Theoretical Guarantee (Main Theorem A: Directional Safety of FN)

Assumptions.

A1 (Geometry). All contrastive representations are ℓ2-normalized; sim(·,·)is cosine similarity, so "moving away from the true centroid" is well-defined on the unit sphere.

A2 (FN weighting and controllability). Potential FNs in the denominator enter with effective weight $w_{FN}=g\cdot\pi_{FN}$, where $\pi_{FN}$is the soft mask (Eq. (24)) and $g$is an uncertainty-aware gate.

A3 (Observable FN mass ratio). Define the normalized FN softmax mass ratio

$$\theta_i = \frac{\sum_{u\in\mathcal{N}_{FN}(i)} w(u)\exp\left(\mathrm{sim}(z_i,u)/\tau\right)}{\sum_{v\in\mathcal{N}(i)} w(v)\exp\left(\mathrm{sim}(z_i,v)/\tau\right)} \tag{25}$$

where $d_i$denotes the unit direction that increases the angular distance from the true centroid c_{y_i}.

We refer to the **FN-safe region** as the set of (g, π_{FN}, τ) such that $\theta_i \le \Theta_i$(with $\Theta_i$estimable from batch statistics), which ensures the FN-induced update has non-positive projection onto the divergent direction $d_i$away from the true centroid, i.e., $\langle \nabla_{z_i}\mathcal{L}, d_i\rangle \le 0$. This makes the theorem's "directional safety" guarantee directly testable and explicitly controlled by $\pi_{FN}$and g.

Main Theorem.

Using the FN softmax mass ratio $\theta_i$defined above, Appendix (Main Theorem A) shows that under mild assumptions and controllable hyperparameters (τ, π_{FN}, g), if $\theta_i \le \Theta_i$ (where Θ_i is a batch-statistics-estimable threshold constructed from the positive gradient term and cross terms), then the projection of a single update onto the divergent direction away from the true centroid is non-positive, i.e., $\langle \nabla_{z_i}\mathcal{L}, d_i\rangle \le 0$. Therefore, the anchor will not be pushed away from its correct cluster, explaining how Eq. (24) mitigates FN-induced disruption in early training. We prove in **Appendix Main Theorem A** that under weak assumptions and controllable hyperparameters (τ, FN masking coefficient, g), when $\theta_i \le \Theta_i$ (where Θ_i is the threshold that can be approximated by batch statistics, encompassing the numerator positive example gradients and cross terms), a single update's projection along the "divergent direction from the cluster center" will be non-positive, meaning that the anchor will not be pushed away from the correct cluster. This result explains the role of Equation (24) in mitigating the "incorrect negative sample disruption" during the early stages of training. The proof and threshold construction are detailed in Appendix Main Theorem A. Thus, FN samples contribute to the negative contrast with only 10% weight, effectively mitigating the disruption of intra-cluster compactness caused by strong repulsion while suppressing the amplification of label errors; Compared to directly masking FN samples, this soft weighting avoids training instability caused by gradient discontinuities and prevents the loss of intra-cluster information due to the complete disregard of positive sample signals. In our implementation, the soft mask is instantiated as $\pi_{FN} = 0.1$ to down-weight potential FNs; the theoretical analysis keeps $\pi_{FN} \in (0,1)$ to represent a tunable masking strength, with the effective FN contribution controlled by $w_{FN} = g \cdot \pi_{FN}$. If i is classified as HN, its feature representation z_i is used to identify two key reference points within the set of cluster centers $\{u_c^{(v)}\}_{c=1}^{C}$ the first is the same-cluster center $u^+ = u_{y_i}^{(v)}$, where y_i represents the cluster assignment index of the sample in the current iteration's clustering division, corresponding to its same-cluster center; the second is the nearest different-cluster center u^-. Based on these two reference points, a push-pull contrastive loss for HN samples is constructed.

$$\mathcal{L}_{\text{push}}^{(v)}(i) = max\{0, \cos\left(z_i^{(v)}, u^+\right) - \cos\left(z_i^{(v)}, u^-\right) + m\}, \tag{26}$$

$$L_{\text{pull}}^{(v)} = 1 - \cos\left(z_i^{(v)}, u^+\right) \tag{27}$$

$$\mathcal{L}_{\text{push}} = \frac{1}{|H|}\sum_{i\in H} \mathcal{L}_{\text{push}}^{(v)}(i), \mathcal{L}_{\text{pull}} = \frac{1}{|H|}\sum_{i\in H} \mathcal{L}_{\text{pull}}^{(v)}(i) \tag{28}$$

Here, H represents the set of samples classified as HN in the current batch, and $m > 0$ is the cosine margin parameter. The first term explicitly enlarges inter-cluster discrimination by increasing the similarity between the sample and its assigned (same-cluster) center u^{+}, while decreasing its similarity to the nearest different-cluster center u^{-} in angular space. The second term continuously enhances the directionality alignment between the sample and the true different-cluster as long as the sample has not sufficiently moved away from the cluster. These two terms are linearly combined with a weight ratio $\lambda_{\text{push}}:\lambda_{\text{pull}}$, and after being multiplied by the batch gating coefficient, the result is the cross-view aggregated hard-negative branch loss.

$$g = \text{clip}\left(\frac{\bar{u} - \mu_{\text{start}}}{\mu_{\text{end}} - \mu_{\text{start}}}, 0,1\right), \mathcal{L}_{\text{HN}} = g\left(\lambda_{\text{push}}\mathcal{L}_{\text{push}} + \lambda_{\text{pull}}\mathcal{L}_{\text{pull}}\right) \tag{29}$$

Here, $\mathcal{H}$ represents the set of samples identified as HN in the current batch, and $\bar{u}$ denotes the average uncertainty of the batch. μ_{start} and μ_{end} are the penalty activation threshold and saturation value, respectively. This design ensures that when $\bar{u} < \mu_{\text{start}}$ (indicating model convergence) or $\bar{u} > \mu_{\text{end}}$ (indicating excessive noise), $\text{g} = 0$,effectively shielding the penalty for pushing difficult negative samples away and preventing interference with the main gradient. In the range between these two values, g increases linearly with $\bar{u}$, allowing the push-away strength to adapt smoothly and appropriately to the difficulty of the samples.

In practice, $\theta_i$can be computed directly from mini-batch softmax weights by summing the (masked-and-gated) FN terms in the InfoNCE denominator and normalizing by the full denominator mass. We estimate $\Theta_i$using batch statistics as described in Appendix Main Theorem A, and then adjust $g$and/or π_{FN}(or equivalently the effective FN weight w_{FN}) to keep $\theta_i \leq \Theta_i$, which makes the directional-safety condition verifiable during training.

3.6 Imbalanced Sample Mitigation (ISM)

To address the issue of insufficient samples in minority clusters, the ISM module introduces pseudo-sample compensation during feature learning. It ensures the effectiveness of these compensated samples through distribution alignment and repulsion regularization, and ultimately reinforces the clustering signals of minority clusters by incorporating a cluster-level contrastive loss.

3.6.1 Minority Cluster Detection and Pseudo-Sample Generation

On the consensus feature set $\{z_i^{\text{cons}}\}_{i=1}^{N}$, we have obtained the cluster labels $l_i \in \{1, \dots, C\}$ and the corresponding cluster centers u_c^{cons} for each sample using K-MEANS, as detailed in Section 3.3. Additionally, we compute the sample count for each cluster, $n_c = |\{i: l_i = c\}|$, and calculate the average cluster size $\overline{n} = \frac{1}{C}\sum_{j=1}^{C} n_j$ Based on a predefined global minority cluster ratio threshold $\rho \in (0,1)$, we define the global minority cluster set $S = \{c \mid n_c < \rho\overline{n}\}$.

In the current training batch, let$\tilde{L}$ be the set of cluster indices that appear, and the minority cluster indices requiring compensation are defined by $M = S \cap \tilde{L}$. For each $c \in M$, if the sample count in the batch $n_c \leq 1$, the center is directly reverted to the initial center $\psi_c = \mu_c$; otherwise, pseudo-samples are generated according to the following steps.

First, calculate the mean of the samples in this cluster:

$$\mu_c = \frac{1}{n_c}\sum_{i:l_i=c} z_i^{\text{cons}} \tag{30}$$

As well as the covariance.

$$\Sigma_c = \frac{1}{n_c - 1 + \varepsilon} \sum_{i:l_i = c} (z_i^{\text{cons}} - \mu_c)(z_i^{\text{cons}} - \mu_c)^{\top}, \tag{31}$$

Where $\varepsilon > 0$ is a numerical stability term. To introduce greater diversity in the feature space, the covariance is further amplified by the number of pseudo-samples, P.

$$\Sigma_c' = \frac{P}{n_c} \Sigma_c + \varepsilon I_D \tag{32}$$

By reparameterization sampling, we generate pseudo-samples.

$$\psi_{c,p} \sim N(\mu_c, \Sigma_c'), p = 1, \ldots, P. \tag{33}$$

Here, I_D is $D \times D$ identity matrix, and Σ_c' must be positive definite.

3.6.2 Distribution Alignment and Exclusion Regularization

After generating pseudo-samples, we apply the maximum mean discrepancy alignment and exclusion regularization for each $c \in M$. Specifically, we use a Gaussian kernel with bandwidth $\sigma > 0$.

$$K(z_i, \psi_{c,p}) = \exp\left(-\frac{\| z_i - \psi_{c,p} \|^2}{2\sigma^2}\right) \tag{34}$$

The MMD within the batch is estimated as

$$D_{\text{MMD}}^{(c)} = \frac{1}{n_c(n_c - 1)} \sum_{i,j=1 i \neq j}^{n_c} K(z_i, z_j) + \frac{1}{P(P-1)} \sum_{p,q=1 p \neq q}^{P} K(\psi_{c,p}, \psi_{c,q}) - \frac{2}{n_c P} \sum_{i=1}^{n_c} \sum_{p=1}^{P} K(z_i, \psi_{c,p}) \tag{35}$$

Here, z_i represents the representation of the i-th real sample within a cluster in the latent space, and the pseudo-samples are indexed by $p = 1..P$ The alignment of the distributions between pseudo-samples and real samples is achieved by minimizing $\sum_{c \in M} \lambda_{\text{mmd}} \mathrm{D}_{MMD}^{(c)}$.

To prevent the pseudo-samples of different minority clusters from aggregating, let the exclusion bandwidth $\sigma_r > 0$. $c, c' \in M$ represent two distinct clusters in the minority cluster set M. The exclusion loss is defined as

$$\mathcal{L}_{\text{rep}} = \sum_{\substack{c,c' \in M \\ c \neq c'}} \sum_{p=1}^{P} \sum_{q=1}^{P} \exp\left(-\frac{\| \psi_{c,p} - \psi_{c',q} \|^2}{2\sigma_r^2}\right) \tag{36}$$

And minimize it to encourage the pseudo-samples to move away from each other. Additionally, impose boundary constraints on each pseudo-sample.

$$\mathcal{L}_b = \sum_{c \in M} \sum_{p=1}^{P} \left[max(\| \psi_{c,p} - u_c^{cons} \| - R_{max}, 0) + max(R_{min}' - \min_{j \neq c} \| \psi_{c,p} - u_j^{cons} \|, 0) \right], \tag{37}$$

Where $R_{\max}$ and $R'_{\min}$ represent the maximum radius of the current cluster and the minimum distance threshold to the centers of other clusters, respectively. By weighting and combining the three components mentioned above, the overall imbalance regularization loss is obtained.

$$L_{\mathrm{imb}} = \sum_{k \in M} \left(\lambda_{\mathrm{mmd}} \mathrm{D}_{MMD}^{(c)} \right) + \lambda_{\mathrm{rep}} L_{\mathrm{rep}} + \lambda_{\mathrm{b}} L_{\mathrm{b}}, \tag{38}$$

Where $\lambda_{\mathrm{mmd}}, \lambda_{\mathrm{rep}}, \lambda_{\mathrm{b}} \geq 0$ are adjustable hyperparameters.

Theoretical guarantee (Main Theorem B: stable centroids of minority clusters and approximate margin preservation)

Assumptions.

B1 (Cluster separability). Let $\gamma > 0$denote a minimal inter-center gap (between the minority center and other class centers).

B2 (Intra-cluster pseudo-sampling boundary). Let $R_{\max}$be an upper bound on the radius of minority pseudo-samples around the minority center, and let $R_{\min}$be a lower bound on the distance from the minority center to other centers. We require $R_{\max} < \gamma/2$ and $R_{\min} > \gamma/2$.

These boundary conditions ensure pseudo-samples remain in an intra-cluster "safe region," so prototype/centroid updates reduce variance without inducing cross-cluster drift, consistent with the boundary constraint mechanism used in the algorithm.

Main Theorem.

The ISM implementation in this section employs Gaussian sampling and distribution alignment (see Eqs. (32)-(35)), combined with boundary constraints (Eq. (37)). We prove in **Appendix Main Theorem B** that, under mild batch independence and a minimal inter-cluster gap $\gamma > 0$, if a minority cluster c generates pseudo-samples via the current Gaussian scaling scheme (Eqs. (32)-(33)), then, after merging P pseudo-samples with n_c real samples, the new centroid estimate $\hat{\mu}_c^{\mathrm{new}}$ satisfies a trace-wise variance reduction:

$$\mathrm{tr}(\mathrm{Var}(\hat{\mu}_c^{\mathrm{new}})) \leq \frac{1}{n_c + P} \mathrm{tr}(\Sigma_c') \leq \frac{1}{1 + P/n_c} \cdot \frac{1}{n_c} \mathrm{tr}(\Sigma_c) + O(\varepsilon),$$

that is, the overall (non-directional) variance of the centroid estimator is suppressed by a factor of $\frac{1}{1+P/n_c}$; when $P \approx n_c$, the variance is roughly halved. With more general elliptical/sub-Gaussian sampling and optional covariance shrinkage, the theorem remains valid (the current implementation corresponds to the special case of "no shrinkage"; see the appendix for details).

Meanwhile, let $\sigma_{max} = \sqrt{\lambda_{max}(\Sigma_c')}$ and define

$$R_{max} = \sigma_{max}(\sqrt{D} + t) + \hat{\delta}, R_{min}' = \gamma - R_{max},$$

where t is chosen according to the target default rate $\bar{\alpha}$ and satisfies a sub-Gaussian tail bound. Then, with probability $\geq 1 - 2Pe^{-c_0 t^2}$, all pseudo-samples lie within the cluster radius R_{max}, and their distance to the nearest non-matching centroid is at least R_{min}'; together with the repulsive and boundary regularization in Eqs. (36)-(37), this ensures approximate margin preservation and no boundary violations. This result explains the source of ISM's effect of "stable centroid and preserved margin" on minority clusters; see Appendix Main Theorem C for the proof and parameter-setting details.

After completing pseudo-sample generation and regularization for minority clusters, we incorporate $\{\psi_{c,p}\}$ into the cluster-level contrastive loss to simultaneously strengthen positive signals and maintain negative boundaries. The consensus centroids are $\{u_c^{cons}\}_{c=1}^{C}$, with temperature $\tau > 0$. For cluster c, its contrastive loss consists of two parts: First is positive enhancement. For each centroid u_c^{cons}, the original principal positive is its match with the target centroid, $\exp\,(\frac{\cos\,((\mu_c^{cons},\mu_c^{(v)})}{\tau})$; when cluster c is a minority cluster, we additionally introduce positives from its generated pseudo-samples, $\exp\left(\frac{\cos(u_c^{cons},\psi_c)}{\tau}\right)$. Therefore, the total positive score of cluster c is

$$P_c = \exp\left(\frac{\cos\,(\mu_c^{cons},\mu_c^{(v)})}{\tau}\right) + \mathbf{1}[c \in M]\cdot\exp\left(\frac{\cos\,(\mu_c^{cons},\psi_{c,p})}{\tau}\right) \tag{39}$$

In this way, minority clusters retain the centroid's "principal positive" while gaining auxiliary positives from pseudo-samples, markedly reinforcing their intra-cluster aggregation signal.

The negative score is composed of other cluster centroids $\{\mu_j^{cons} \mid j \neq c\}$ and pseudo-samples from other minority clusters $\{\psi_{j,p} \mid j \in \mathcal{M}, j \neq c, p = 1, \dots, P\}$:

$$\begin{aligned} N_c^{\text{real}} &= \sum_{j\neq c} \exp\left(\frac{\cos\left(u_c^{cons},u_j^{cons}\right)}{\tau}\right) \\ N_c^{\text{pseudo}} &= \sum_{\substack{j\neq c\\ j\in M}} \exp\left(\frac{\cos\left(u_c^{cons},\psi_{j,p}\right)}{\tau}\right) \end{aligned} \tag{40}$$

and they are merged into the cluster-level total negative score

$$N_c = N_c^{\text{real}} + N_c^{\text{pseudo}} \tag{41}$$

Finally, the contrastive loss of cluster c is

$$\ell_c = -\log\frac{\mathrm{P}_c}{\mathrm{P}_c + \mathrm{N}_c + \epsilon} \tag{42}$$

and we average over all clusters

$$L_{\mathrm{c}}^{(v)} = \frac{1}{C}\sum_{c=1}^{C} \ell_c \tag{43}$$

In this process, the positives of minority clusters include both the real centroid and pseudo-samples, while the negatives encompass other centroids and pseudo-samples, thereby achieving the dual goals of positive reinforcement and negative-boundary maintenance. This strategy enables minority clusters in contrastive learning to obtain sufficient attraction signals without pseudo-samples eroding inter-cluster margins.

3.7 Three-Stage Training and Overall Objective Loss

Let the current training epoch be denoted as e, the warm-up phase threshold as E_w, and the cross-view introduction phase threshold as E_c. To achieve a progressive curriculum learning approach, the entire training process is divided into three stages, with the corresponding loss terms being dynamically introduced at each stage. The final objective loss is the combination of the stage-specific losses and the imbalance regularization.

3.7.1 Stage 1: Cluster Structure Warm-Up

When the training epoch e satisfies e $\leq E_w$, the model undergoes a warm-up phase using three types of loss functions: cluster-level contrastive loss, sample-level contrastive loss, and uncertainty regression loss.

Sample-level contrastive learning (SCL) maximizes the similarity between same-cluster samples and suppresses the similarity between negative sample pairs by calculating the cosine similarity between a sample in view v and the global consensus representation. This enhances feature consistency across different views. Let the positive example set for anchor i be $P(i) = \{j: y_j = y_i\}$ (using pseudo-labels), with weights $w_{ij} = y_{ij}^{\text{pse}}$. Let $r \in N_i^{(v)}$ where $N_i^{(v)}$ represents the set of negative samples for sample i that do not belong to the same cluster, including samples from both the current view v and the global consensus space v_c, and $N_i^{(v)} = \{r \mid W_{ir} = 1, y_r^{pse} \neq y_i^{pse}\}$, then:

$$\mathcal{L}_{\text{feat}}^{(v)} = \frac{1}{N}\sum_{i=1}^{N}\left[-\log\frac{\sum_{j\in P(i)} w_{ij}\exp\left(\cos\left(z_i^{(v)}, z_j^{\text{cons}}\right)/\tau\right)}{\sum_{r=1}^{N}\exp\left(\cos\left(z_i^{(v)}, z_r^{(v)}\right)/\tau\right)}\right] \tag{44}$$

The uncertainty regression loss $L_u^{(v)}$ minimizes the difference between the model-predicted uncertainty u_i and the true uncertainty u_i calculated from the membership distribution. The specific formula is:

$$L_u^{(v)} = \frac{1}{N}\sum_{i=1}^{N}(\hat{u}_i - u_i)^2 \tag{45}$$

By minimizing this loss, the model adapts to the uncertainty in the data during training, providing a reliable foundation for the subsequent discrimination of false-negative and hard-negative samples.

The loss function for Stage 1, $L^{(1)}$, is defined as:

$$L^{(1)} = \sum_{v=1}^{V}\left[\alpha L_{\text{c}}^{(v)} + \beta L_{\text{feat}}^{(v)} + \lambda_u L_u^{(v)}\right], \tag{46}$$

where λ_u represents the weight of the uncertainty regression loss.

Through the synergistic optimization of these loss terms, the model in Stage 1 gradually adjusts and stabilizes its feature representations, thereby providing a more stable and discriminative foundation for the subsequent training phases.

3.7.2 Stage Two: FN/HN Subnet Classification and Penalty

When the training epoch reaches the stage $E_w < e \leq E_c$, based on the cluster centers and uncertainty statistics converged in Stage 1, a push-pull penalty is introduced for hard-negative samples. The corresponding loss is:

$$L^{(2)} = L^{(1)} + \sum_{v=1}^{V}\lambda_{\text{hn}}g\left(L_{\text{push}}^{(v)} + L_{\text{pull}}^{(v)}\right) \tag{47}$$

$\lambda_{\text{hn}} > 0$ is the weight coefficient of the hard-negative sample push-pull penalty term, used to balance its contribution to the overall loss. Additionally, the MLP serves solely as a soft router, without introducing extra supervision, and its parameters are updated end-to-end via the chain gradients of $\mathcal{L}_{\text{HN}}$, $\mathcal{L}_{\text{feat}}^{(v)}$, and y^{pse}. Moreover, during the early stages, the gradients of the bottom 10% of samples with the lowest confidence are halted to enhance stability.

Theoretical Guarantee (Main Theorem C: Angular Separation Growth Across Two Time Scales)

Assumptions.

C1 (Two-time-scale dynamics). Embeddings are updated on a fast time scale (learning rate α), while prototypes/centroids are updated more slowly (e.g., every Tsteps or with $\beta \ll \alpha$).
C2 (Active-margin condition). The margin constraint is active (e.g., $s^+ - s^- + m > 0$) so that the push-pull mechanism applies.
C3 (Stable step size). The embedding step size satisfies a stability upper bound implied by the theorem, under which the angular separation between positives and negatives increases monotonically as training progresses.

Main Theorem.
In Stage 2, the "center is updated every T steps" (corresponding to the cycle t in K-Means, where $T = t$) and a push-pull penalty is applied to the HN samples (Equations (26) - (29)). In Appendix, **Main Theorem C** proves that under spherical geometry and two-time-scale dynamics, if the current activation $s^+ - s^- + m > 0$ (cosine separation trigger), there exists an explicit upper bound for the learning rate such that the angular separation between the anchor and the correct cluster center strictly increases, and the lower bound for the pull term's contribution remains non-negative. The result is framed probabilistically to characterize the event of "correctly selecting the nearest other cluster." This conclusion ensures that HN samples will not remain trapped at the boundary for extended periods.

To improve readability, we provide a summary table that links our theoretical guarantees to the corresponding modules and their controlling hyperparameters (Table 2). This mapping also serves as a quick reference for the hyperparameter roles in the subsequent implementation and ablation analyses.

Table 2. Mapping from theorems to modules and controlling hyperparameters

Theorem	Mechanism	Key hyperparameters	What is controlled
Theorem A - Directional safety of FNs	Soft FN masking + uncertainty-aware gating in weighted contrastive denominator	π_{FN}(soft mask; cf. Eq. (24)), g(gate), τ(temperature)	FN softmax mass ratio θ_i is upper-bounded to prevent push-away along the divergent direction from the true centroid ($\langle \nabla_{z_i} \mathcal{L}, d_i \rangle \leq 0$)
Theorem B - Minority-centroid stability	Minority pseudo-sampling + boundary constraint + variance-reduced centroid update	(inter-center gap), $R_{\max}$, $R_{\min}$, pseudo-sampling scale parameters	Enforces $R_{\max} < \gamma/2$ and $R_{\min} > \gamma/2$ so pseudo-samples stay intra-cluster, reducing centroid variance without cross-cluster drift
Theorem C - Monotonic angular-margin enlargement	Margin-based push-pull with two-time-scale optimization (fast embeddings, slow prototypes)	m(margin), α(embedding LR), β or T(prototype update rate/frequency)	Under active margin and stable step size, angular separation between positives and negatives increases monotonically across training

3.7.3 Stage Three: Cross-View Consistency Enhancement

Once the training epoch $e > E_c$, and as the pseudo-labels, cluster centers, membership degrees, and uncertainty estimates gradually stabilize, the model enters the later stages of training. In this phase, we introduce the cross-view weighted InfoNCE loss to further enhance the consistency between the multi-view representations and the consensus representation. Unlike traditional InfoNCE, which relies on label indicator functions, we adaptively construct the weighted relationships between samples through an uncertainty module.

Specifically, for the batch sample set $\{x_i\}_{i=1}^{N}$, we first calculate the consistency score between samples based on their membership vectors $\mu_i = [\mu_{i1}, \dots, \mu_{iL}]$ and pseudo-labels y_i. The consistency score S_{ij} is defined as: $S_{ij} = 1_{[y_i=y_j]} \cdot min(\mu_{i,y_i}, \mu_{j,y_j})$. This matrix element S_{ij} represents the soft co-membership confidence level that sample i and sample j belong to the same cluster, with the value derived from the combination of their membership degrees and pseudo-labels. Subsequently, each row of S is normalized to obtain a probability distribution in the form of weights.

$$\hat{S}_{ij} = \frac{S_{ij}}{\sum_{l \neq i} S_{il}}, j \neq i \tag{48}$$

This provides each anchor sample i with a soft label distribution $\hat{S}_{ij}$ as probabilistic cluster assignments concerning the potential positive sample set.

Building on this, for the representation matrix $Z^{(v)} = [z_1^{(v)}, \dots, z_N^{(v)}]$ of the v-th view, $z_r^{(v)}$ represents the feature representation of any sample r in the batch under view v. The weighted InfoNCE loss is defined as:

$$\mathcal{L}_{\text{wNCE}}(Z^{(v)}; \hat{S}, \tau) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left[-\sum_{j \neq i} \hat{S}_{ij} \log \frac{\exp\left(\cos\left(z_i^{(v)}, z_j^{(v)}\right)/\tau\right)}{\sum_{r \neq i} \exp\left(\cos\left(z_i^{(v)}, z_r^{(v)}\right)/\tau\right)} \right] \tag{49}$$

Here, $\mathcal{I} = \{i: \sum_{l \neq i} S_{il} > 0\}$ represents the set of valid anchor points. This loss function essentially minimizes the cross-entropy between the soft label distribution $\hat{S}_{ij}$ determined by the uncertainty module and the softmax distribution based on similarity, thereby introducing a confidence-weighted flexible alignment constraint between the views.

Furthermore, we not only compute the weighted InfoNCE for each individual view as described above but also extend it to the consensus representation $Z^{cons} = [z_1^{(c)}, \dots, z_N^{(c)}]$, thus obtaining the overall cross-view consistency loss:

$$\mathcal{L}_{\text{cross}} = \sum_{v=1}^{V} \mathcal{L}_{\text{wNCE}}(Z^{(v)}; \hat{S}, \tau) + \mathcal{L}_{\text{wNCE}}(Z^{cons}; \hat{S}, \tau) \tag{50}$$

Through this design, the cross-view weighted InfoNCE not only mitigates pseudo-label noise by utilizing soft labels but also enhances the semantic consistency between multi-view representations and the consensus space through uncertainty-driven adaptive weights, thereby significantly improving the model's clustering performance and the discriminative power of its representations.

Finally, this loss is adjusted within the overall training objective using a gating factor g and view-weighting coefficient β_{p}. The complete total loss at this stage is given by:

$$L^{(3)} = L^{(2)} + \beta_{\text{p}} \cdot g \cdot L_{\text{cross}}^{(v)} \tag{51}$$

Here, β_{p} is the view-weighting coefficient, which controls the overall contribution of the cross-view consistency loss, while g is the gating factor dynamically adjusted according to the training progress. It increases with the training epochs to prevent early unstable signals from adversely affecting the model's convergence.

Finally, the class imbalance regularization term L_{imb} is added to the third-stage loss, resulting in the overall optimization objective for the entire network:

$$L_{\text{total}} = L_{\text{imb}} + L^{(3)} \tag{52}$$

Through the above three-stage curriculum learning design, the model can achieve preliminary convergence under the most straightforward and reliable cluster-level and feature consistency signals, then specifically apply fine-grained penalties to difficult negative samples. Finally, based on stable pseudo-labels and uncertainty estimates, the model fully leverages the complementary advantages of multi-view representations, ultimately achieving a robust contrastive learning objective with compact intra-cluster and well-separated inter-cluster structures.

This approach enables the model to quickly establish a basic structure of "intra-cluster cohesion and inter-cluster dispersion" under the initial clustering centers. Our model's training process is outlined in Algorithm 1.

3.8 Complexity Analysis

The time complexity of the model primarily arises from the following aspects: the computations involved in deep network propagation, the operations within the self-supervised learning module, and the loss calculations within the same module. First, during the forward propagation of the network, the time complexity is approximately $O(VN^2D + VN^2\log\ k + VN^2)$, where k denotes the number of neighbors in the FN/HN submodule. Second, in the contrastive learning module, the computation of similarities between node embeddings and contrastive losses incurs a complexity of $O\big(VN^2D + ENVD + E'ND(k + B)\big)$, where B represents the batch size and E the number of pretraining epochs. Finally, the loss calculation within the self-supervised module has a complexity of $O(max\{N^2, B^2, C^2, n^2, m^2\}D)$, with m denoting the number of pseudo-samples. Taken together, the overall time complexity of the model can be expressed as: $O(VN^2D + ENVD + E'ND(k + B) + max\{N^2, B^2, C^2, n^2, m^2\}D)$

Algorithm 1 shows the procedure of the MVSIB.

```
Input: Multi-view data $X = \{x_i^{(v)}\}_{i=1,v=1}^{N,V}$; $C$; Hyper-parameters
Pre-training: Compute the sum of reconstruction losses over all views and update the autoencoder parameters using this loss.
Extract the consensus representation $z^{cons}$ and initialize the cluster centers $u_c^{(v)}$
Set $\rho_e \leftarrow \rho_0$ (by Eq. (13)); Instantiate an Adam optimizer with weight decay.
For $e = 1\ to$ E do
    Update the pseudo-labels $y_i$ using the current $z^{(v_c)}$
    For each batch do
        Compute per-view latent embeddings $z^{(v)}$ and $z^{cons}$;
        Compute $m_{ic}^{(v)}$ an $u_i^{(v)}$ (by Eqs. (7) and (11));
        Partition the "certain" set $C_t$ and "uncertain" set $U_t$ according to $\rho_t$ (by Eq. (13);
        For each $i \in U_t$ , compute f $_i$ ,predict $\hat{y}_i$ via an MLP (by Eqs. (18) and (22));
        Identify M and sample $\psi_{c,p}$ ; augment these into $L_c$ (by Eqs. (33) - (43));
        Apply distribution alignment and repulsion regularization (by Eqs. (36) and (37));
        Compute $L_{\text{total}}$ and update (by Eq. (52));
        Backpropagate and update all model parameters.
    end for
end for
Output: Return the final clustering labels $y_i$ obtained by hardening the learned soft assignments for $z^{cons}$and the cluster centers $u_c^{(v)}$.
```

4. Experiments

4.1 Benchmark Datasets

In this experiment, we selected multiple multi-view clustering datasets to evaluate the proposed algorithm, specifically including Prokaryotic [44], RGB-D [45], CORA [46], CCV [47], and Hdigit [48]. Detailed information for each dataset can be found in Table 3.

Table 3. Characteristics of Benchmark Datasets Used in MVSIB Evaluation

Datasets	Samples	Classes	Views	Dimensionality (Different views)
Prokaryotic	551	4	3	438/3/393
RGB-D	1449	13	2	2048/300
Cora	2708	7	4	2708/1433/2708/2708
CCV	6773	20	3	5000/5000/4000
Hdigit	10000	10	2	784/256

Prokaryotic: This dataset, provided by NCBI, consists of 551 Prokaryotic samples and includes three types of feature views. View 1 is the textual feature view, represented by a bag-of-words model of species documents, reflecting semantic attributes. View 2 is the proteomic composition view, characterized by the relative frequency of amino acids to represent molecular composition features. View 3 is the gene family feature view, constructed based on the presence or absence of gene families, reflecting genomic characteristics.

RGB-D (Sentences NYU v2): This dataset is derived from the NYU Depth v2 and contains 1,449 indoor scene images, divided into 13 categories. Its features consist of two views: the first view is a 2048-dimensional image feature extracted by a ResNet-50 model pre-trained on ImageNet, representing visual information; the second view is a 300-dimensional text embedding generated by doc2vec, pre-trained on Wikipedia corpus, representing semantic information.

Cora: This dataset was publicly released by the CiteSeer/LINQS project group, containing 2,708 paper nodes and 5,429 citation edges, categorized into 7 classes. Each node provides feature information from four perspectives: the content view, which represents the paper's text using a 1,433-dimensional bag-of-words vector; the in-link view, which reflects how the paper is cited by other papers; the out-link view, which shows how the paper cites other papers; and the integrated view, which combines both in-link and out-link relationships to form a paper-citation network structure.

CCV (Columbia Consumer Video): This dataset, provided by Columbia University, primarily consists of video samples sourced from user-generated content on online video platforms such as YouTube, covering a variety of real-world scenes and activities. The dataset contains a total of 6,773 video samples, categorized into 20 classes. Each sample is represented by a Bag-of-Words based on handcrafted features, forming three views: SIFT, STIP, and MFCC.

Hdigit: This dataset contains 10,000 handwritten digit samples, each composed of features from two perspectives. The two perspectives correspond to image representations from the MNIST and USPS handwritten digit datasets, offering feature information from different sources and styles. Hdigit is often regarded as a bi-view extension of

MNIST and USPS, and is widely used in MVC research.

4.2 Evaluation Metrics

To assess the clustering performance of the model, this study employs five commonly used metrics: Accuracy (ACC) [49], Normalized Mutual Information (NMI) [50], Purity (PUR) [51], Adjusted Rand Index (ARI) [52], and Fowlkes-Mallows Score (FM) [53]. Among them, ACC measures the consistency between the clustering results and the true labels; NMI evaluates the normalized mutual information between the clustering results and the true labels; PUR represents the proportion of the dominant class in each cluster; ARI measures the similarity between the clustering results and the true labels while correcting for the effect of random partitioning; and FM is the geometric mean of pairwise precision and pairwise recall, used to characterize the consistency between the clustering and the true partition in terms of sample pairs. The higher the values of these metrics, the better the performance of the model.

4.3 Comparison methods

We compared MVSIB with nine mainstream and state-of-the-art MVC methods on five publicly available classic multi-view datasets.

K-means [54]: A classic clustering method that partitions data by minimizing the distance between data points and cluster centers.

DEMVC (2021) [55]: Combines multi-view autoencoders with cross-view consistency learning. By co-optimizing deep embeddings and clustering structures, it enhances the expressiveness of multi-view data. It employs an adaptive view selection mechanism to effectively reduce the interference from irrelevant views, improving the model's robustness to noise and enhancing the stability of clustering results.

MvAGC (2021) [56]: Integrates multi-view adaptive graph convolutional networks, dynamically adjusting the weight of adjacency matrices across views to achieve effective information fusion. This model uses adaptive graph convolution layers to capture both local and global structural features across views, significantly improving MVC performance.

MFLVC (2022) [57]: Based on multi-view feature learning and view coordination mechanisms, it jointly optimizes feature representations and view weights to enhance the complementarity between views. The design of a joint low-rank constraint effectively captures the intrinsic structure of the data, aiding in improved clustering accuracy.

DealMVC (2023) [58]: Adopts a deep consistency enhancement strategy, combining multi-view autoencoders with consistency loss to strengthen the expression of shared information across views. Its multi-level feature fusion mechanism promotes coordinated updates of view features, improving the robustness and accuracy of clustering.

GCFAggMVC (2023) [59]: Based on graph convolutional networks and feature aggregation strategies, it dynamically fuses multi-view adjacency matrices and feature information. The attention mechanism designed for weight adjustment enables effective integration of different view features, thereby improving clustering performance.

DCMVC (2024) [60]: Uses a deep collaborative learning framework, combining self-supervised signals and contrastive learning to enhance consistency across multi-view features. Its dual-encoder structure helps extract high-quality cross-view embeddings, facilitating more accurate clustering divisions.

SCMVC (2024) [61]: Employs sparse coding and consistency constraints combined with shared representation learning for inter-view information. By jointly optimizing reconstruction errors and consistency losses, the model effectively suppresses noise impacts, enhancing the stability and accuracy of MVC.

ACCMVC (2025) [62]: A weighted clustering method based on contrastive learning, utilizing a dynamic weight adjustment mechanism and contrastive constraints on multi-view features to enhance semantic consistency across views. This method improves the discriminative power of the embedding space and clustering quality through contrastive loss optimization.

4.4 Implementation Details

In the preparation phase of model training, we selected fully connected layers (Fc) as the core building blocks of the deep network architecture and uniformly configured them. First, we performed pre-training on both the encoder and decoder for 200 epochs with the goal of minimizing the reconstruction error of each view's input as much as possible. Once the main training phase began, the model underwent joint optimization based on multi-view features, with the training period for this stage set to 100 epochs. Throughout the entire training process, the batch size remained fixed at 256, the learning rate was set to 0.0001, the weight decay was set to 0, the encoder's output dimension was 256, and the temperature parameter was fixed at 0.5. During each training epoch, the model dynamically updated pseudo-labels, cluster centers, and uncertainty estimates, subsequently classifying samples based on uncertainty to identify false-negative and hard-negative samples, which were then optimized together with various loss terms.When the sample size is fewer than 10,000, we build the nearest neighbor graph across the entire dataset and select the nearest neighbor number from $\{3, 5, 7, 10\}$. For sample sizes greater than 10,000, the nearest neighbor graph is constructed on small batches of data, selecting the nearest neighbor number from $\{2, 3, 4, 5\}$. Additionally, hyperparameters λ_u and λ_{hn} are chosen from the set $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. To ensure reproducibility of the experimental results, all experiments were conducted using a random seed. The model was run in an environment equipped with an Intel i5-1240P CPU, NVIDIA GTX 3070 GPU, and 32 GB of RAM, and deployed on a Windows 10 system based on PyTorch 3.9.0.

4.5 Clustering Results

The average clustering results after 100 epochs of training are shown in Table 4. MVSIB significantly outperforms nine mainstream and state-of-the-art MVC methods in terms of ACC, NMI, Purity, ARI, and FM across five publicly available datasets (RGB-D, Prokaryotic, Cora, CCV, and Hdigit). By analyzing their design ideas and experimental performance, we can delve into two key advantages of our method:

(1) MVC Methods Without Contrastive Learning

When comparing K-means, MvAGC, and DEMVC, three MVC methods that do not incorporate a contrastive learning mechanism, it is evident that their clustering metrics are consistently lower than those of MVSIB across different tasks. On the RGB-D dataset, the ACC of K-means, MvAGC, and DEMVC are 40.44%, 40.44%, and 35.52%, respectively, while MVSIB achieves 52.94%. On the Prokaryotic dataset, K-means has an ACC of 49.48%, and DEMVC has an ACC of 47.80%, while MVSIB reaches 57.35%. On the Hdigit dataset, K-means, MvAGC, and DEMVC achieve ACCs of 52.91%, 57.00%, and 41.48%, respectively, whereas MVSIB significantly improves to 99.11%. K-means relies solely on simple feature concatenation and struggles to extract meaningful information in

Table ~~3~~4. Average Clustering Metric for MVSIB and Baselines over Five Public Datasets

		Shallow MVC		Deep MVC				Contrastive-based Deep MVC			**Ours**
Dataset	Metric	K-means	DEMVC	MvAGC	ACCMVC	SCMVC	GCFAgg	MFLVC	DealMVC	DCMVC	**MVSIB**
RGB-D	ACC	40.44±0.0653	35.52±0.0348	40.44±0.0246	25.37±0.0232	33.68±0.0136	25.37±0.0232	38.30±0.0575	30.65±0.0583	47.44±0.0267	**52.94±0.0571**
	NMI	37.93±0.0522	19.43±0.0555	32.94±0.0166	22.78±0.0311	28.40±0.0143	22.78±0.0311	20.30±0.0404	15.19±0.0882	40.34±0.0234	**42.70±0.0522**
	PUR	53.28±0.0724	35.52±0.0348	50.79±0.0246	42.87±0.0245	49.84±0.0036	42.87±0.0245	42.31±0.0366	33.77±0.0785	59.80±0.0239	**62.11±0.0472**
	ARI	21.82±0.0483	12.64±0.0431	21.16±0.0140	11.43±0.0214	26.95±0.0110	11.43±0.0214	19.78±0.0584	08.54±0.0778	29.24±0.0210	**35.73±0.0599**
	FM	15.23±0.0613	30.30±0.0491	30.42±0.0299	21.37±0.0183	07.51±0.0084	21.37±0.0183	31.80±0.0517	20.39±0.0144	37.94±0.0196	**43.86±0.0547**
CCV	ACC	19.92±0.0425	11.00±0.0072	15.72±0.0030	29.67±0.0206	27.04±0.0752	28.41±0.0254	26.87±0.0175	22.09±0.0356	35.85±0.0186	**37.87±0.0232**
	NMI	17.79±0.0314	02.93±0.0567	11.38±0.0011	25.28±0.0175	27.57±0.0400	23.46±0.0265	22.01±0.0179	22.18±0.0633	33.20±0.0137	**33.32±0.0174**
	PUR	22.57±0.0506	11.00±0.0072	14.57±0.0078	32.98±0.0194	31.24±0.0983	31.32±0.0256	29.46±0.0176	23.40±0.0393	39.50±0.0146	**40.56±0.0195**
	ARI	06.45±0.0255	01.20±0.0014	02.86±0.0062	12.64±0.0124	12.81±0.0190	11.31±0.0186	15.07±0.0098	09.48±0.0314	18.21±0.0128	**19.04±0.0172**
	FM	06.59±0.0385	20.85±0.0292	10.85±0.0086	17.39±0.0117	17.59±0.0519	16.12±0.0177	10.19±0.0107	18.40±0.0069	22.90±0.0127	**23.54±0.0163**
Cora	ACC	36.74±0.0384	31.55±0.0162	27.48±0.0244	28.60±0.0182	40.41±0.0397	25.47±0.0139	32.92±0.0368	35.05±0.0206	61.46±0.0050	**64.55±0.0276**
	NMI	15.11±0.0283	15.44±0.0117	05.81±0.0255	13.45±0.0367	26.11±0.0050	07.93±0.0161	17.77±0.0394	22.01±0.0105	**45.86±0.0114**	44.28±0.0296
	PUR	38.18±0.0455	31.55±0.0162	01.14±0.0244	36.90±0.0269	48.07±0.0204	34.70±0.0192	42.28±0.0316	45.76±0.0016	**65.95±0.0068**	65.93±0.0206
	ARI	02.85±0.0224	07.30±0.0160	28.36±0.0209	07.37±0.0217	16.60±0.0197	04.41±0.0124	10.74±0.0385	11.41±0.0014	34.65±0.0046	**39.83±0.0266**
	FM	16.83±0.0354	28.49±0.0259	30.83±0.0299	22.90±0.0141	30.66±0.0014	20.31±0.0131	26.24±0.0244	28.14±0.0029	46.38±0.0045	**50.25±0.0233**
Hdigit	ACC	52.91±0.0752	41.48±0.0618	57.00±0.0512	96.60±0.0101	-	97.02±0.0069	83.45±0.1404	75.64±0.0928	99.00±0.0135	**99.11±0.0257**
	NMI	47.17±0.0683	34.80±0.0777	49.14±0.0407	91.33±0.0178	-	92.19±0.0126	72.91±0.1738	83.42±0.0620	97.33±0.0218	**97.57±0.0343**
	PUR	54.73±0.0803	41.48±0.0618	38.09±0.0481	96.60±0.0101	-	97.02±0.0069	75.27±0.1506	77.86±0.0831	99.00±0.0135	**99.11±0.0257**
	ARI	32.99±0.0653	23.61±0.0851	58.95±0.0437	92.65±0.0207	-	93.52±0.0143	83.94±0.1344	71.29±0.1014	97.86±0.0251	**98.06±0.0427**
	FM	49.63±0.0732	36.69±0.0853	44.49±0.0427	93.38±0.0186	-	94.17±0.0129	75.67±0.1556	69.92±0.0342	98.07±0.0226	**98.26±0.0384**
Prokaryotic	ACC	49.48±0.0454	47.80±0.0742	-	52.44±0.0589	55.90±0.026	53.07±0.0141	52.78±0.0612	56.44±0.0268	51.87±0.0388	**57.35±0.0556**
	NMI	10.22±0.0953	06.52±0.0604	-	26.75±0.0253	22.83±0.019	26.59±0.0131	11.90±0.1115	04.85±0.0841	26.49±0.0239	**27.38±0.0249**
	PUR	50.32±0.0542	47.80±0.0742	-	60.68±0.0223	63.33±0.016	61.86±0.0088	57.97±0.0244	57.49±0.0321	**66.53±0.0257**	64.34±0.0250
	ARI	10.53±0.0442	02.48±0.0505	-	14.85±0.0361	12.54±0.014	**18.51±0.0122**	05.88±0.0642	03.31±0.0576	16.72±0.0247	18.47±0.0394
	FM	29.34±0.1063	49.29±0.0623	-	43.04±0.0230	45.07±0.013	44.71±0.0072	52.95±0.1030	**60.26±0.0720**	52.12±0.0675	58.07±0.0596

Our method's average clustering performance (%) compared with 9 other comparison methods.

Bold values are the best results, underlined values are the second best results; “-” indicates that the method is not mentioned.

the presence of noisy views. MvAGC's graph-based regularization fails to fully compensate for semantic features when minority class samples are sparse. DEMVC, while aligning view label distributions, is prone to introducing cross-view biases. In contrast, MVSIB generates pseudo-samples for imbalanced clusters and aligns distributions at the cluster level. This not only strengthens the representation of underrepresented clusters but also significantly enhances the model's ability to distinguish between easily confused samples. As a result, MVSIB demonstrates superior clustering performance compared to these three contrast-free methods.

(2) MVC Methods with Contrastive Learning

Among the mainstream contrastive learning methods such as ACCMVC, GCFAggMVC, SCMVC, DealMVC, DCMVC, and MFLVC, the performance of DCMVC on the Cora dataset serves as an example, with ACC and ARI scores of 61.46% and 34.65%, respectively, and an FM of 46.38%. In comparison, MVSIB achieves 64.55%, 39.83%, and 50.25%, respectively, indicating that relying solely on fixed contrastive objectives struggles to balance view uncertainty and hard-negative samples. On the CCV dataset, DCMVC's NMI and PUR are 33.20% and 39.50%, while MVSIB reaches 33.32% and 40.56%.

In terms of ACC on the Prokaryotic dataset, ACCMVC, DealMVC, and DCMVC achieve 52.44%, 56.44%, and 51.87%, while MVSIB reaches 57.35%. For the FM metric on Hdigit, MFLVC, DealMVC, and DCMVC report FM values of 75.67%, 69.92%, and 98.07%, while MVSIB achieves 98.26%. This demonstrates that models relying on fixed contrastive targets often struggle to address view uncertainty and class-imbalanced cluster samples. The rigid selection of positive and negative samples leads to the omission of hard-negative samples and lacks the adaptability to compensate for intra-cluster distribution bias. MVSIB, through uncertainty-guided false-negative and hard-negative sample selection, curriculum-based negative sample mining, and dynamic contrastive weights, combined with pseudo-sample completion and cluster-level distribution alignment, preserves the cross-view discriminative advantages of contrastive loss while achieving adaptive compensation in class-imbalanced scenarios. This results in robust and leading clustering performance across various complex heterogeneous tasks.

4.6 Parameter analysis

In this section, we analyze the impact of four key parameters on the performance of MVSIB, measured by the ACC metric, across five datasets. The experimental results for the following parameters are presented in subplots A-D of Figure 6: the number of pre-warming epochs in curriculum learning (E_c), the weight of the uncertainty regression loss (λ_u), the initial coefficient of uncertain sample ratio (p_0), and the entropy-separation balance coefficient (α).

4.6.1. Number of Pre-warming Epochs in Curriculum Learning (E_c)

Subplot A illustrates the variation in ACC across different datasets with varying numbers of pre-warming epochs (E_c). From the figure, it can be observed that as the number of pre-warming epochs increases, the ACC values for RGB-D, CCV, and Prokaryotic datasets gradually decline from initially high levels, with the best performance occurring at 20 epochs. This suggests that excessive pre-warming may lead to unnecessary computational overhead, particularly for simpler or less noisy datasets, where longer pre-warming does not necessarily yield better results. In contrast, CORA and Hdigit datasets maintain stability as the number of pre-warming epochs increases, reaching optimal performance at 40 epochs. This indicates that these datasets are less dependent on the number of pre-warming epochs and do not require extensive pre-warming to achieve effective learning outcomes.

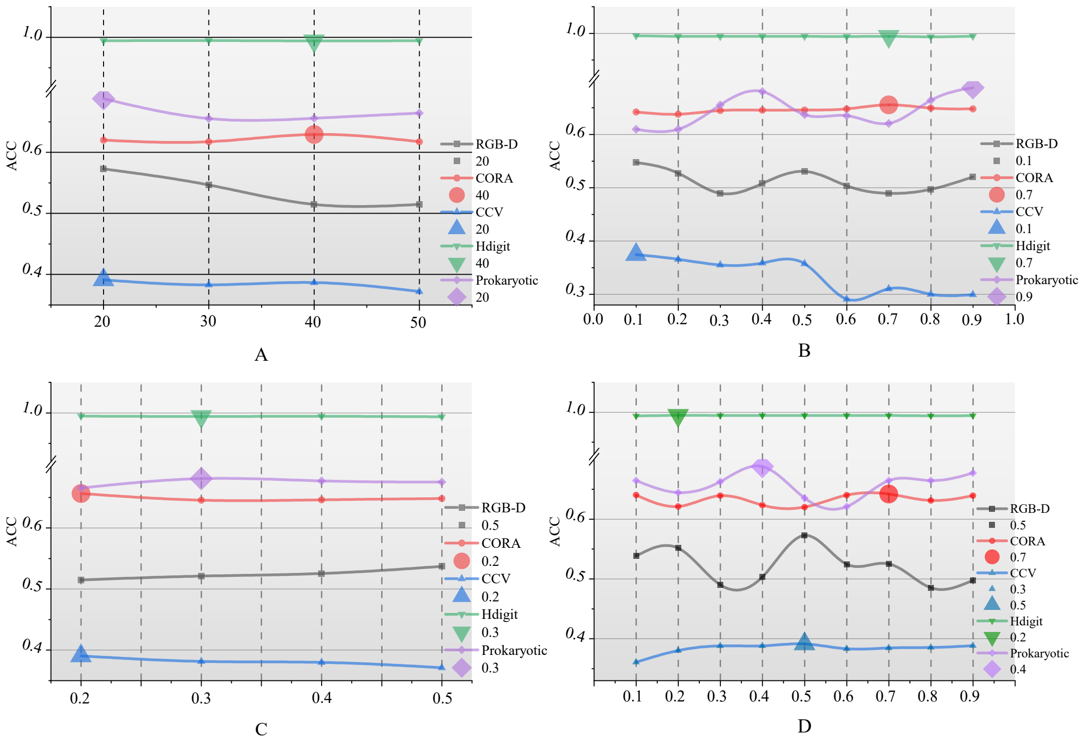


Fig. 6. Hyperparameter Sensitivity Analysis of MVSIB

4.6.2. Weight of Uncertainty Regression Loss (λ_u)

Subplot B presents the variation in ACC across different datasets with different values of the uncertainty regression loss weight (λ_u). λ_u has significant differences in its impact across different datasets. For RGB-D and CCV, the highest accuracy is achieved when λ_u=0.1; as λ_u increases, the RGB-D curve fluctuates, while CCV exhibits a monotonous decline. This suggests that excessively large regression weights may cause the model to overly focus on high-uncertainty samples, thereby weakening the representation of the primary contrastive signals. Prokaryotic shows a more pronounced dependence on λ_u, with a peak at λ_u = 0.9, indicating that the false-negative and hard-negative sample structure in this dataset is more complex, and larger uncertainty regularization helps recalibrate the boundaries. In contrast, the ACC-λ_u curves for CORA and Hdigit remain largely flat, with only a slight advantage at λ_u=0.7, suggesting that their cluster divisions are already well-defined and that they are not sensitive to changes in this weight. Overall, the proper setting of λ_u is particularly crucial for datasets with complex uncertainty structures, while for datasets with higher cluster separation, the model can maintain stable performance across a relatively wide range of λ_u values.

4.6.3. Initial Coefficient of Uncertain Sample Ratio (p_0)

Subfigure C illustrates the variation in ACC across five benchmark datasets at different initial uncertain sample ratios,

p_0. As shown in the graph, the ACC values for all datasets remain relatively stable across various p_0 values, with an overall steady performance. The ACC for the RGB-D dataset reaches its maximum at p_0 = 0.5, exhibiting a slight upward trend, which may suggest that at this ratio, the model achieves a reasonable balance in the proportion of uncertain samples, leading to improved learning performance. For the CCV and CORA datasets, the optimal performance is observed at p_0=0.2, followed by a decline, indicating that for these datasets, a higher initial uncertainty ratio introduces too many uncertain samples, diluting the signal from certain samples. In the case of Hdigit and Prokaryotic, the best performance is observed at p_0=0.3, suggesting that a small adjustment in p_0 primarily affects the early gradient of uncertain samples, without significantly altering the final decision boundary.

4.6.4. Entropy-Separation Degree Balance Coefficient (α)

Subfigure D presents the clustering accuracy across five datasets for different α settings, revealing significant variations across the datasets. The ACC values for the RGB-D and Prokaryotic datasets fluctuate greatly, reaching their optimal points at α=0.5 and α=0.4, respectively, with performance significantly declining just slightly away from the optimal values. This suggests that in scenarios with high uncertainty or severe cluster overlap, the relative weight of the entropy and separation terms determines the decision boundary in the feature space, moderately increasing the entropy weight can alleviate inter-cluster confusion, while excessive weight suppresses the boundary risk term, leading to misclassification of hard-negative samples. In contrast, the curves for CCV, CORA, and Hdigit are smoother, with optimal performance observed at α=0.5, 0.7, and 0.2, respectively, and limited neighborhood fluctuations. This indicates that the samples in these datasets are more stable, and adjustments to α have a smaller impact on model performance, likely because these datasets contain less noise, allowing the model to achieve good performance within a narrower range of parameter adjustments.

4.7 Ablation Analysis

In this section, we conduct a quantitative evaluation of the independent contributions of each module within the proposed model, specifically including: the class imbalance handling module, the FN and HN sample discrimination module, the sample-level contrastive learning module, and curriculum learning. To this end, we successively disable or replace these components on five public MVC datasets: CCV, Hdigit, Cora, RGB-D, and Prokaryotic and then perform a comprehensive comparative analysis of clustering performance under each configuration using multiple metrics, including Silhouette Score, ACC, ARI, NMI, Purity, and FM. This enables us to uncover the performance gains and underlying mechanisms of each module across different data distributions and view structures.

4.7.1. Presence or Absence of a Class Imbalance Module

We conducted a comparison of clustering performance on the RGB-D dataset with and without the class imbalance module, as illustrated in Figure 7. The horizontal axis denotes the cluster ID, while the vertical axis represents the Silhouette Score, a metric used to evaluate clustering quality by reflecting the compactness of samples within clusters and the separation between clusters. For each sample, the Silhouette Score s is defined as $s = \frac{b-a}{\max(a,b)}$.

where a is the average distance between the sample and other samples within the same cluster, and b is the average distance between the sample and those in the nearest neighboring cluster. The value of s lies within [−1, 1]; the closer it is to 1, the better the clustering, indicating that intra-cluster samples are more compact and inter-cluster separation more distinct, whereas values near 0 imply ambiguous clustering results. The blue bars denote the results obtained with the class imbalance module, whereas the gray bars correspond to those without it; the depth of each bar's color signifies the number of samples contained in the cluster. Overall, the average Silhouette Score improved from 0.47 to 0.53, with all clusters showing higher scores compared to the case without the module. Notably, in clusters with fewer samples (such as clusters 10, 11, and 12), the class imbalance module demonstrated a remarkable enhancement. This indicates that the class imbalance module plays a significant role in improving clustering performance, particularly by mitigating the decline in clustering quality caused by insufficient samples in minority clusters, thereby substantially strengthening the algorithm's adaptability in handling imbalanced data.

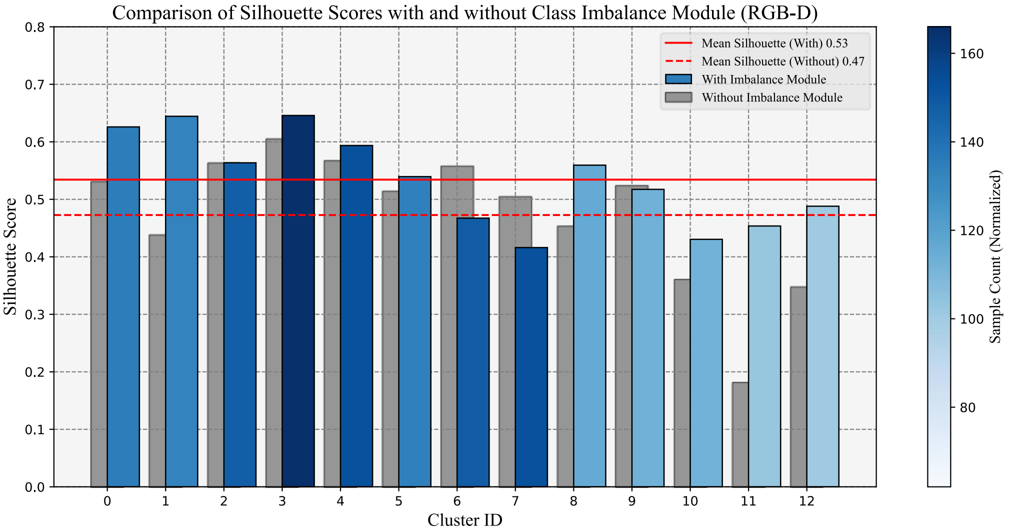


Fig. 7. Impact of Imbalanced Sample Mitigation (ISM) on Silhouette Scores per Cluster

4.7.2. Presence or Absence of Hard Negative and False Negative Handling

Figure 8 illustrates the comparison of five clustering metrics across the RGB-D, CCV, Cora, Hdigit, and Prokaryotic datasets, with and without the incorporation of the false negative and hard negative discrimination module. It can be observed that, after enabling the module, the ACC on the RGB-D dataset increased from 0.42 to 0.58; the CCV dataset exhibited an improvement of approximately 0.25 in PUR, alongside gains of 0.08-0.12 in other metrics; the Cora dataset achieved an average enhancement of around 0.10 across all metrics; the Hdigit dataset rose from 0.80 to 0.99; while the Prokaryotic dataset showed gains of 0.15-0.20 in ACC, PUR, and FM, and improvements in NMI and ARI from about 0.30 to above 0.50. These results demonstrate that false negative discrimination effectively preserves intra-class consistency (PUR, ARI), whereas hard negative discrimination reinforces inter-class separation (ACC, FM). The discrimination module relies on three criteria, differences in cross-view neighborhood distances, similarity between samples and cluster centers, and deviations in cross-view feature distributions, in order to dynamically distinguish false negatives from hard negatives. It then applies soft-weight attenuation to the former and push-pull constraints to the latter, thereby enhancing overall clustering quality.

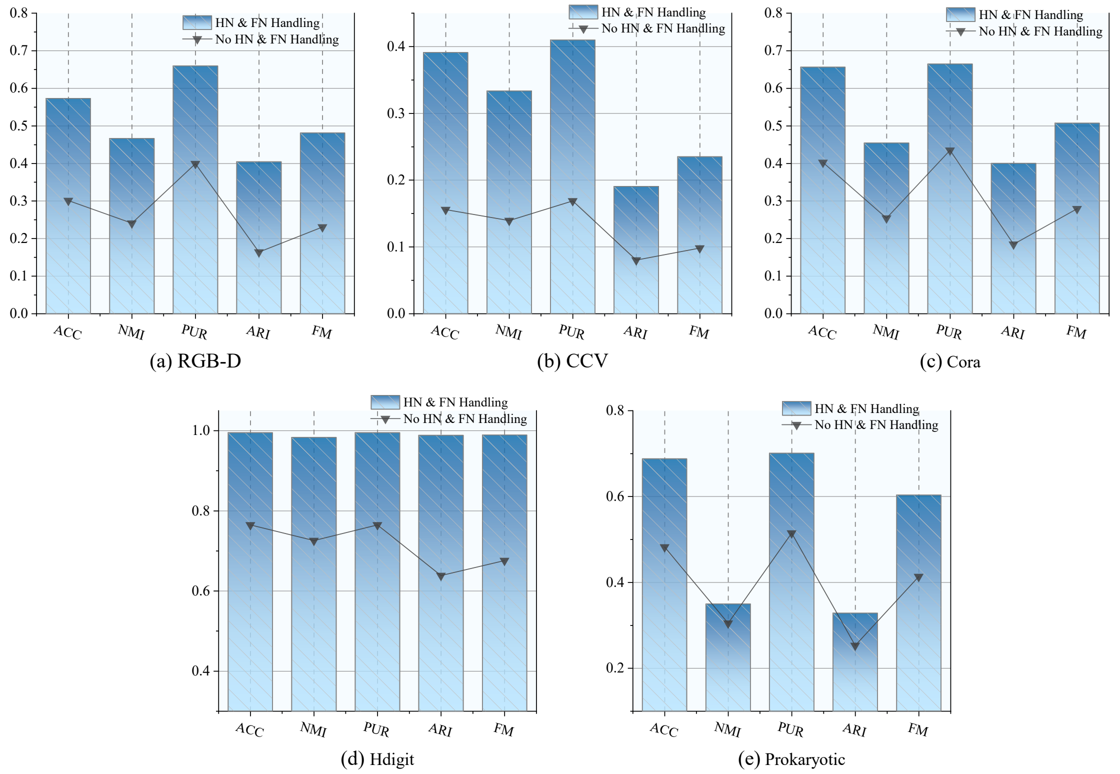


Fig. 8. Ablation Study on FN/HN Disambiguation Module

4.7.3. Presence or Absence of Sample-Level Contrastive Learning

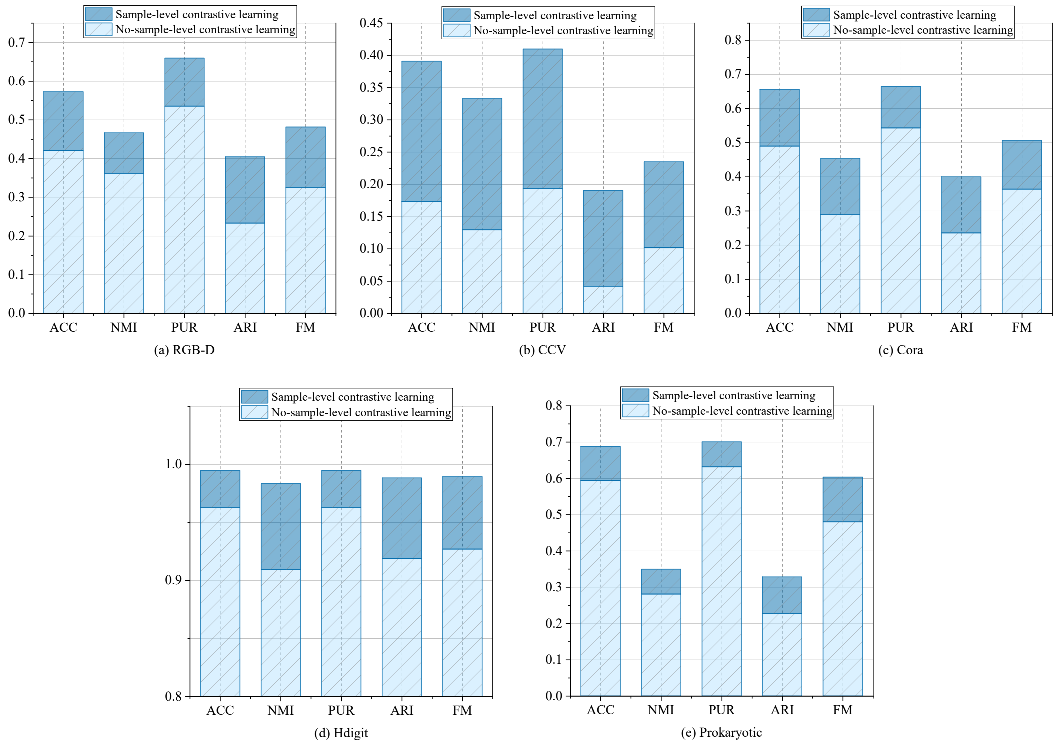


Fig. 9. Ablation of Sample-Level Contrastive Learning (SCL)

As illustrated in Figure 9, we evaluated the impact of incorporating the sample-level contrastive learning module on model performance across five datasets. On each dataset, the inclusion of sample-level contrastive learning significantly enhanced the results. On the RGB-D and CCV datasets, ACC increased by approximately 0.20, with PUR and FM also showing improvements. On the Cora dataset, ACC rose from 0.50 to 0.70, while PUR and FM each improved by around 0.20. For the Hdigit and Prokaryotic datasets, ACC increased by about 0.10 to 0.20, with FM and other metrics likewise exhibiting gains. This improvement can be attributed to the positive and negative similarity constraints imposed on sample pairs, aligning features across different views and thereby ensuring consistency between the consensus graph and the original view in the node representation space. Such a design not only preserves structural information but also effectively alleviates feature bias caused by view discrepancies. Moreover, by constructing soft label masks and similarity matrices for positive and negative samples, the module further mitigates the effects of false negatives and outliers, thereby enhancing both the overall performance and generalization ability of the model. In contrast, models lacking curriculum learning struggle to achieve the same level of effectiveness.

4.7.4 Presence or Absence of Curriculum Learning

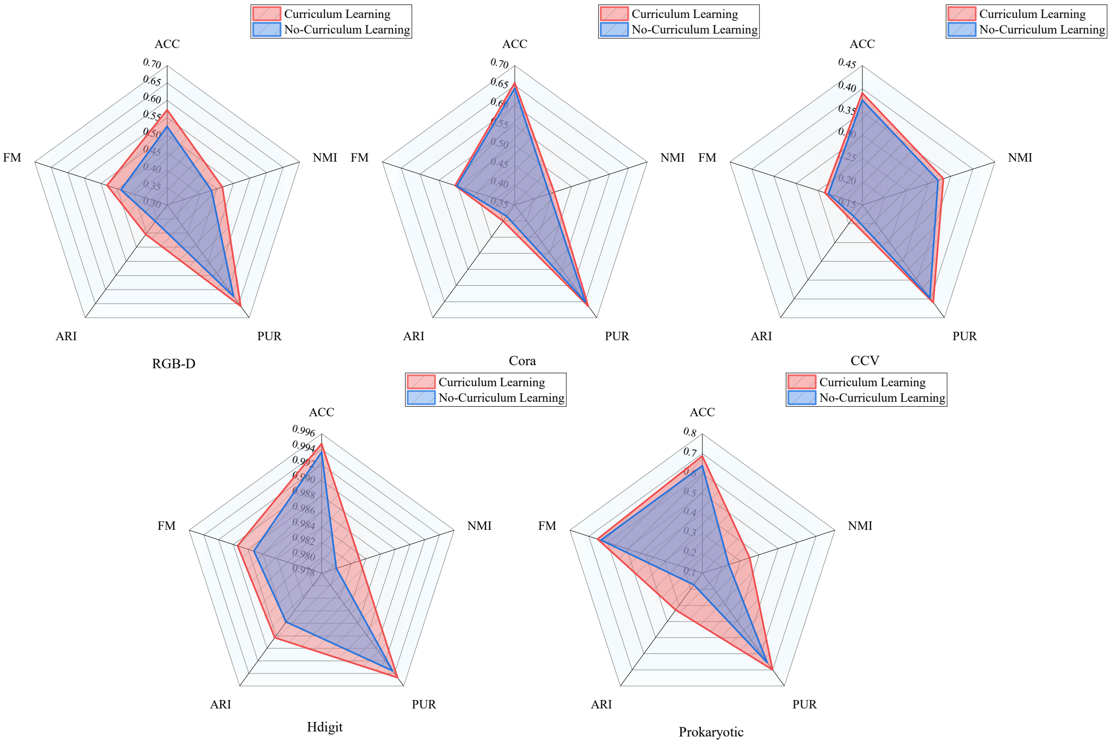


Fig. 10. Curriculum Learning Hyperparameter Analysis

As shown in Figure 10, employing a curriculum learning strategy in our model generally yields superior performance across all datasets compared to training without it. Specifically, on the RGB-D dataset, ACC improved from 0.52 to 0.58, with PUR increasing by about 0.05. On the Cora dataset, ARI rose by approximately 0.01. The Hdigit dataset exhibited gains across ACC, PUR, and FM, with ACC increasing by around 0.0015 and FM by 0.002. On the

Prokaryotic dataset, ACC, PUR, and FM each improved by roughly 0.05, while ARI exceeded 0.3. These results demonstrate that curriculum learning facilitates model convergence by beginning with simpler tasks and gradually advancing to more complex ones, thereby enabling the model to progressively construct a coherent knowledge framework. This approach enhances both learning effectiveness and generalization ability, while avoiding the pitfalls of unsupervised learning, where the complexity of initial clustering tasks and sensitivity to outliers often hinder training. In contrast, models lacking curriculum learning struggle to achieve the same level of effectiveness.

4.8 Statistical Significance Experiments

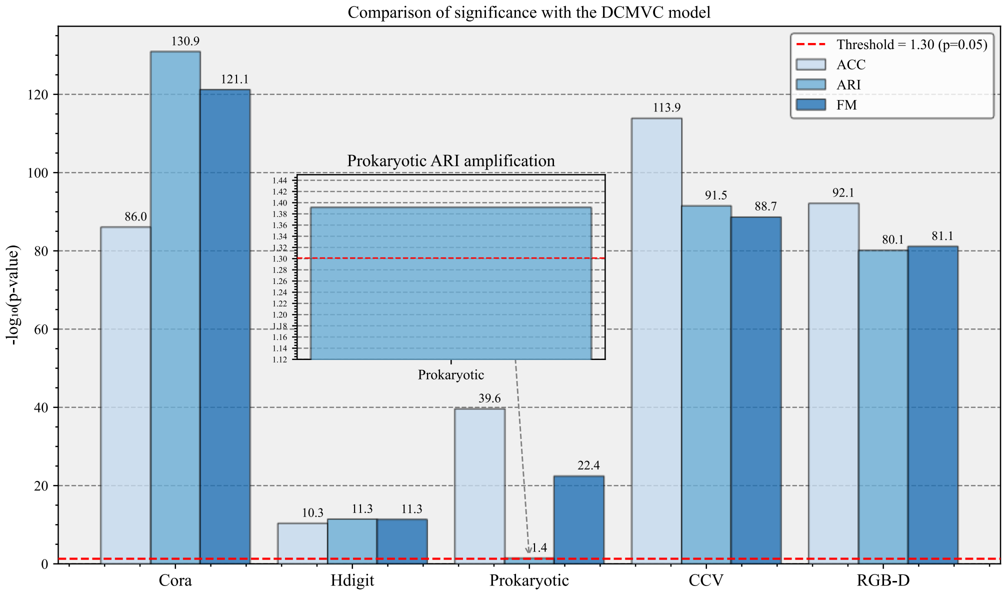


Fig. 11. Statistical Significance of Performance

To verify the statistical significance of the performance improvement achieved by the proposed MVC model, we adopt $-\log_{10}(0.05) \approx 1.30$ as the threshold for significance testing. On five public datasets: CCV, Hdigit, Cora, RGB-D, and Prokaryotic, our comparative algorithm DCMVC serves as the current state-of-the-art baseline. For three key evaluation metrics, ACC, NMI, and FM, we conducted two-sided paired t-tests between MVSIB and DCMVC. As shown in Figure 11, the p-values in all comparative experiments fall well below this threshold, firmly demonstrating that the performance gains of our model are both reliable and systematic rather than accidental outcomes of random fluctuation. This marked improvement is driven by the synergistic effects of multiple modules within our model, which substantially enhance feature discriminability and enable robust and significant performance advantages across diverse data distributions and structures.

4.9 Computational Complexity and Runtime Efficiency

To evaluate computational efficiency, we report the wall-clock execution time of all compared methods on each benchmark dataset, as summarized in Table 5. Figure 12 (a) provides a visual comparison of runtime across datasets.

Table 5. Execution time (in seconds) of different methods on each dataset

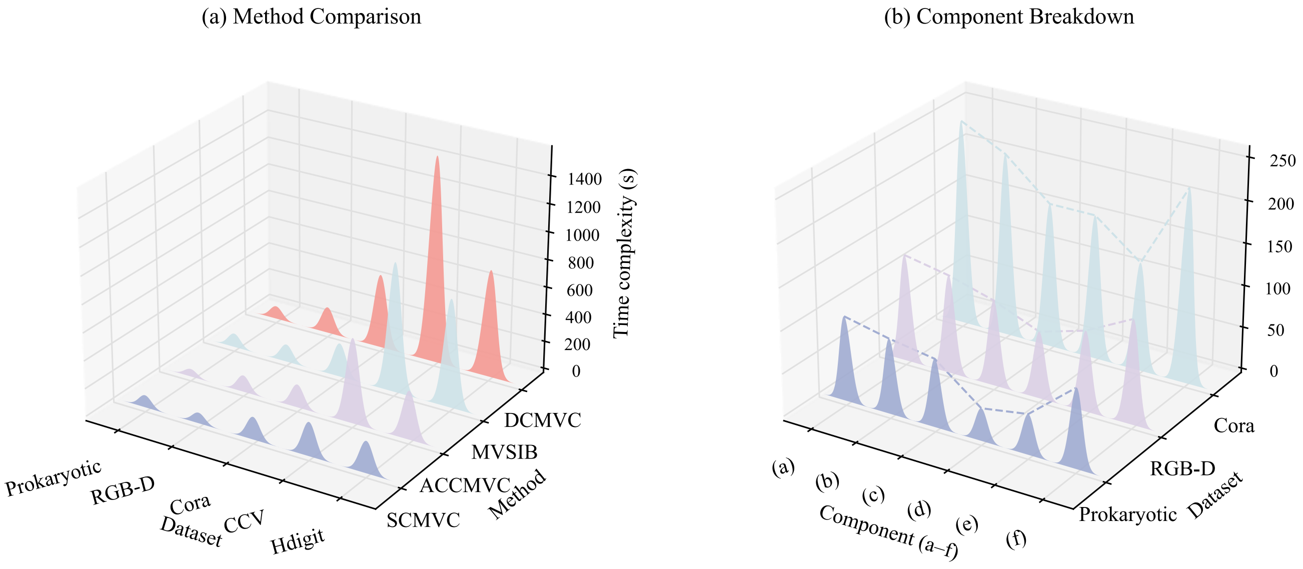
Dataset	SCMVC	ACCMVC	DCMVC	**MVSIB**
Prokaryotic	88.35	55.90	94.33	100.47
RGB-D	76.88	120.18	192.15	127.55
Cora	163.04	167.00	539.71	245.15
CCV	244.08	612.54	1500.16	940.17
Hdigit	228.69	349.42	780.88	783.44

To further identify the major computational bottlenecks within MVSIB, we conduct a component-wise runtime profiling analysis and summarize the results in Table 6 (see also Figure 12 (b)). The component legend is provided in the table note.

Table 6. Execution time (in seconds) of different modules in the ablation study

Dataset	(a)	(b)	(c)	(d)	(e)	(f)
Prokaryotic	100.47	89.04	80.52	38.35	47.45	93.52
RGB-D	127.55	117.20	101.74	79.35	95.33	123.63
Cora	245.15	218.47	172.05	171.74	129.42	230.31

Note: Components are denoted by (a)-(f): (a) Original Model; (b) Pseudo-label Update & Full Forward Pass; (c) KNN Distance in CSD; (d) Pseudo-sample Generation & MMD in ISM; (e) Feature Traversal in Sample-level Contrastive Loss; (f) Build Global KNN Graph.



Component legend: (a) Original Model; (b) Pseudo-label Update & Full Forward Pass;
(c) KNN Distance in CSD; (d) Pseudo-sample Generation & MMD in ISM;
(e) Feature Traversal in Sample-level Contrastive Loss; (f) Build Global KNN Graph.

Fig. 12. Runtime Efficiency Comparison and Component-wise Time Breakdown.

As shown in Figure 12 (a) and Table 5, compared with representative methods from one category: deep MVC (SCMVC) and contrastive-based deep MVC (DCMVC, ACCMVC), MVSIB typically incurs additional runtime overhead due to the extra computations introduced by the proposed modules. However, MVSIB remains substantially more efficient than DCMVC on large-scale datasets. For instance, on CCV, MVSIB takes 940.17 s, which is higher than ACCMVC (612.54 s) but significantly lower than DCMVC (1500.16 s).

The component-level results in Figure 12b and Table 6 further reveal that the main sources of runtime overhead arise from the imbalance-aware operations and kNN-related computations. In particular, the pseudo-sample generation and MMD-based alignment in ISM contribute a notable portion of the overall cost on larger or more complex datasets (e.g., RGB-D and Cora), while kNN distance evaluation and global kNN graph construction also introduce non-negligible overhead. On smaller datasets such as Prokaryotic, the relative cost of MMD-related operations becomes less dominant, and the overall runtime remains comparable across components.

Overall, these results suggest that MVSIB achieves a favorable trade-off between computational cost and clustering performance, making it suitable for applications where improved clustering accuracy and robustness are prioritized.

4.10 Robustness Analysis of the FN/HN Routing Mechanism

This section provides a mechanistic robustness evaluation of the proposed FN/HN routing discrimination. Beyond reporting final clustering scores, we examine whether routing decisions and the associated soft masking and hard repulsion behaviors remain stable and well-controlled when (i) local consistency evidence is corrupted and (ii) view quality becomes heterogeneous. Accordingly, we design two complementary perturbation settings: structural neighborhood noise (Experiment A) and heterogeneous view degradation (Experiment B).

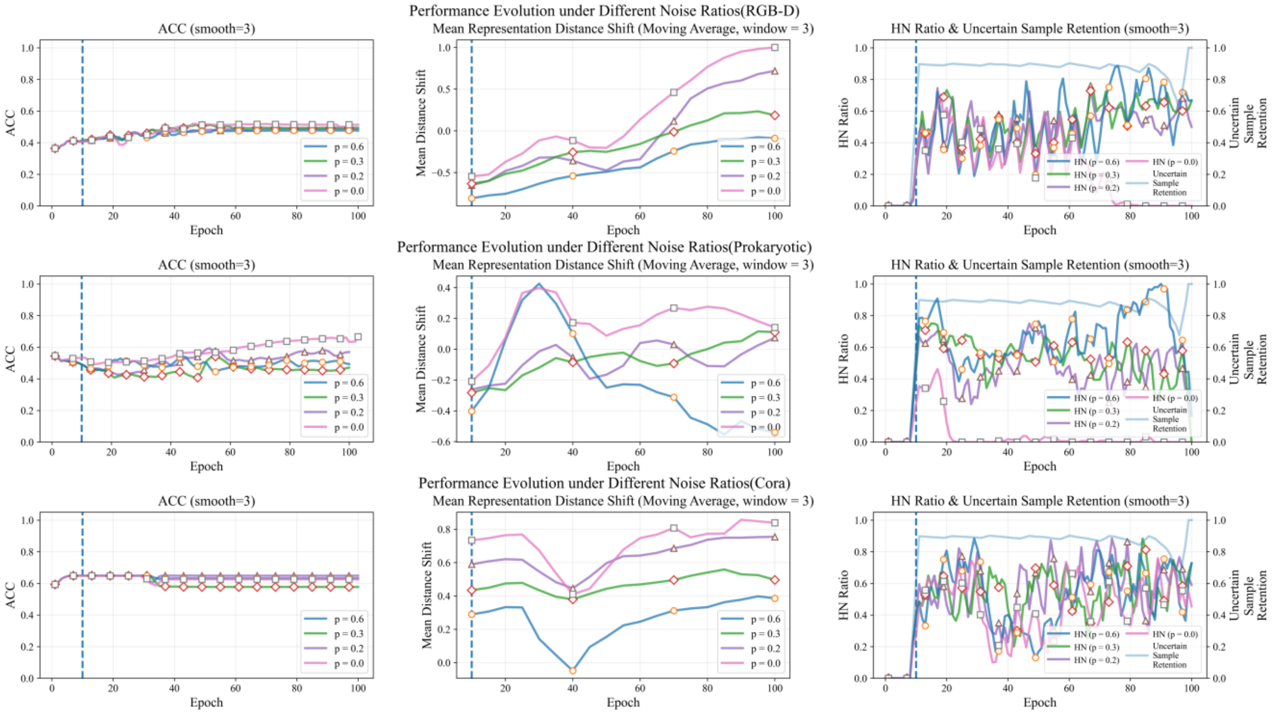


Fig. 13. Robustness of the FN/HN routing mechanism under structural neighborhood noise

4.10.1 Mechanism-level metrics

To directly quantify routing stability, we introduce two mechanism-level metrics. Let $\mathcal{K}_t$ denote the set of samples that pass uncertainty/confidence filtering and participate in routing at epoch t, and let $r_t(i) \in \{\mathrm{FN}, \mathrm{HN}\}$be the routing label. We define Routing Consistency (RC) between adjacent epochs as: $\mathrm{RC}_t = \frac{1}{|\mathcal{K}_t \cap \mathcal{K}_{t-1}|} \sum_{i \in \mathcal{K}_t \cap \mathcal{K}_{t-1}} 1[r_t(i) = r_{t-1}(i)]$. We report the average after routing is activated, denoted as $\mathrm{RC}_{\mathrm{post}}$. A higher $\mathrm{RC}_{\mathrm{post}}$ indicates more temporally stable routing decisions (a moderate decline under stronger perturbation is expected; values approaching the random level ≈0.5 would suggest instability).

We also define the HN Trigger Rate as: $\mathrm{HN}_t = \frac{|\{i \in \mathcal{K}_t : r_t(i) = \mathrm{HN}\}|}{|\mathcal{K}_t|}$. and report the post-activation average $\mathrm{HN}_{\mathrm{post}}$, which captures whether hard repulsion is overly triggered under perturbations.

4.10.2 Experiment A (Structural neighborhood noise)

We apply a targeted structural perturbation to the key input signal of the discriminator. Specifically, when rebuilding the kNN graph at each epoch, we randomly replace $p\%$ of the k nearest neighbors of each sample with random samples ($p \in \{0, 0.2, 0.3, 0.6\}$). This manipulation preserves the original multi-view features while directly disrupting the local voting structure on which Δd depends. Figure 13 reports the evolution of ACC, the mean Δd, HN ratio, and uncertain-sample retention under different noise levels. As p increases, Δd is consistently weakened and becomes persistently low (or even negative) under severe noise, confirming that the perturbation effectively compromises local-consistency evidence. Nevertheless, training remains stable and convergent, with ACC showing smooth degradation rather than collapse. Moreover, although the HN ratio increases and becomes more fluctuating under stronger noise, it does not saturate toward 1, while the uncertain-sample retention remains high, indicating that hard repulsion is not excessively triggered and that the routing behavior remains controlled under structural interference.

4.10.3 Experiment B (Heterogeneous view quality)

We further consider a more realistic scenario in multi-view learning where different views exhibit systematic differences in information quality, reliability, or completeness. To simulate such heterogeneous view-quality conditions, we adopt dataset-adaptive degradations: (i) for RGB-D, we introduce sample-level missingness to mimic depth-view failure; (ii) for Cora, we apply random feature-dimension missingness to model partial semantic loss in high-dimensional sparse features; and (iii) for Prokaryotic, we inject scale/offset perturbations to emulate distribution shifts caused by measurement changes or batch effects. Despite their different implementations, all degradations share the same goal of weakening the reliability of one view in cross-view consistency discrimination, thereby stress-testing the stability of FN/HN routing.

Figure 14 summarizes mechanism-level robustness under heterogeneous view quality, using $\mathrm{RC}_{\mathrm{post}}$ and HN Trigger Rate as two orthogonal axes. Across RGB-D, Cora, and Prokaryotic, most settings remain in the region of high $\mathrm{RC}_{\mathrm{post}}$ and controlled HN triggering even under severe degradations, indicating that routing does not degenerate into unstable or near-random assignment. Notably, $\mathrm{RC}_{\mathrm{post}}$ does not monotonically decrease with degradation intensity; in some cases (e.g., Prokaryotic and Cora) it even increases, suggesting that routing decisions remain temporally consistent when cross-view evidence is weakened. Meanwhile, the HN Trigger Rate does not exhibit explosive growth and may decrease on some datasets, implying that hard repulsion is not excessively activated by low-quality views. Together, these results support that the observed performance drop under high degradation is mainly due to reduced effective cross-view information rather than instability of the FN/HN routing mechanism.

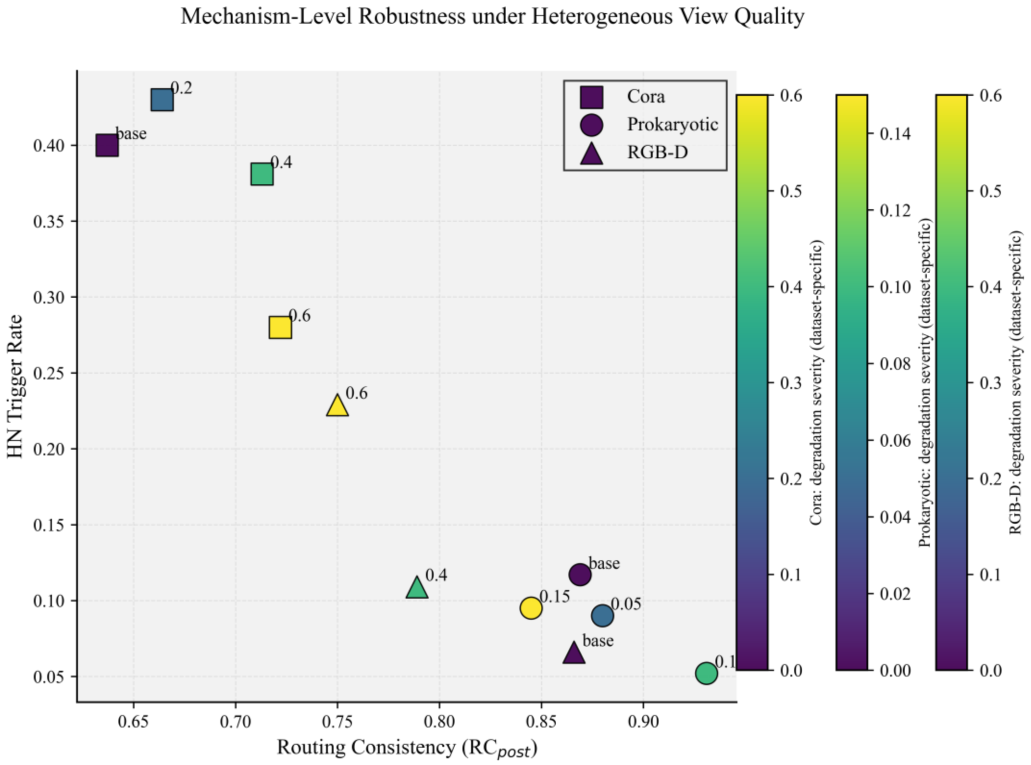
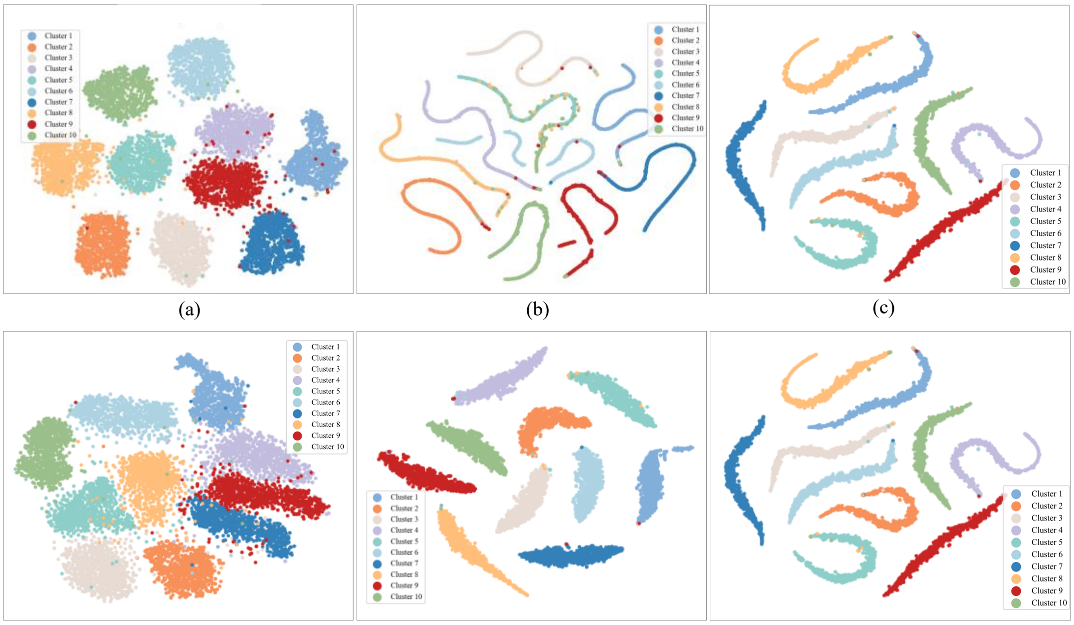


Fig. 14. Mechanism-Level Robustness of FN/HN Routing under Heterogeneous View Quality

4.11 Visualized analysis

In Figure 15, the six subplots of the Hdigit dataset, displayed within the same t-SNE coordinate system, provide an intuitive comparison of clustering quality from the perspectives of feature type and training progression. In the first row: (a) when the raw low-level features of each view are concatenated and projected, the point clouds are heavily intermingled and clusters are indistinguishable; (b) with the introduction of multi-view instance contrastive learning, intra-cluster coherence improves and inter-cluster separation begins to emerge, yet significant overlap persists, with blurred cluster boundaries and instances of internal fragmentation; (c) under the complete MVSIB model, which incorporates adaptive class-imbalance compensation together with false-negative and hard-negative mining, the clusters become highly compact and mutually distinct, with sharply reduced intra-cluster variance and markedly increased inter-cluster separation, thereby validating the superiority of this approach in extracting discriminative consensus representations. The second row illustrates training evolution: (d) depicts the initial distribution of raw data; (e) after 50 iterations, the combination of contrastive loss and dynamically updated pseudo-labels has already induced a preliminary separation of cluster structures; (f) by the 100th iteration, the model converges, with clusters clearly delineated and free from noisy bridging, further demonstrating that the model effectively enhances both the class separability and intra-cluster compactness of view-invariant features, thus laying a solid foundation for subsequent clustering analysis based on the learned embeddings.



(a) (b) (c)

(d) (e) (f)

Fig 15. t-SNE Visualization

5. Discussion

Although MVSIB performs well in clustering on the five datasets: Prokaryotic, RGB-D, Cora, CCV, and Hdigit, it faces a series of intertwined bottlenecks due to its deep coupling of GMM-based pseudo-sample augmentation, MMD distribution alignment and repulsion kernel regularization, comprehensive KNN graph construction, and uncertainty-guided sample selection. On the one hand, the GMM assumption struggles to capture the multimodal complexity in Prokaryotic and the multi-peak structure of features in CCV and RGB-D, resulting in pseudo-samples that often lack sufficient diversity and representativeness, thus failing to effectively reinforce minority cluster features. On the other hand, as the dataset size approaches or exceeds ten thousand, the computational and memory overhead of distribution alignment and repulsion kernel regularization grows superlinearly, leading to a significant surge in memory usage and prolonged training time, thereby limiting deployment on larger-scale visual or biological data sets. Meanwhile, slight adjustments in key hyperparameters, such as the number of warm-up iterations and the uncertainty threshold, can cause large fluctuations in clustering accuracy on RGB-D and Prokaryotic, making the model highly sensitive and dependent on manual hyperparameter tuning when transferring across heterogeneous datasets, which severely impacts reproducibility and plug-and-play functionality. Moreover, the sample selection strategy based on entropy and the Top-2 similarity difference occasionally misclassifies boundary or noisy samples as false-negative/hard-negative, weakening the discriminative power of soft masking and adversarial loss, and reducing inter-cluster separability. To further enhance MVSIB's robustness and scalability in more complex, diverse, and large-scale scenarios, future work could explore more flexible generative models, scalable approximate computation techniques, and dynamic uncertainty metrics based on graph topology.

6. Conclusion

The primary contribution of this work is a modular, collaboratively optimized deep multi-view contrastive learning framework that addresses key MVC challenges, including false-negative misidentification, insufficient hard-negative mining, and class-imbalance-induced performance bottlenecks. First, by employing a consensus-space approach, minority clusters are automatically identified, and pseudo-samples are generated through a Gaussian Mixture Model, thereby enriching class-wise sample distributions. The authenticity and diversity of these pseudo-samples are further ensured via MMD-based distribution alignment and repulsive-kernel regularization. Subsequently, an uncertainty regression coupled with a dynamic weighting mechanism partitions samples into “certain” and “uncertain” classes, imposing differentiated contrastive penalties that effectively mitigate the interference of noisy and hard negative samples. Finally, cross-view weighted InfoNCE leverages consensus consistency scores across views to deeply integrate multi-view information, tightly coupling the objectives of contrastive learning and clustering. Extensive ablation studies and comparative experiments demonstrate that this method achieves consistent and significant performance improvements on the RGB-D, Cora, CCV, Hdigit, and Prokaryotic datasets. Nevertheless, the current GMM-based pseudo-sample generation struggles to fully capture true sample characteristics in multi-modal distributions, while MMD alignment and KNN graph construction remain computationally costly on large-scale datasets. Moreover, the method remains sensitive to several hyperparameters (e.g., the FN/HN routing thresholds and the dynamic-weighting coefficients), which currently require manual tuning and may lead to noticeable performance fluctuations, limiting stability and reproducibility in real-world applications. To address this limitation, future research could introduce adaptive adjustment mechanisms by employing meta-learning or Bayesian optimization techniques to automate hyperparameter selection. Such an approach would enable the model to self-adjust according to dataset-specific characteristics, thereby enhancing robustness, scalability, and generalization across diverse tasks and datasets.

Contribution statement

Binyu Zhao: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing-Original draft, Review & Editing. **Binxiong Li:** Methodology, Software, Validation, Investigation, Data curation, Writing-Original draft, Review & Editing. **Heyang Gao:** Methodology, Software, Validation, Investigation, Data curation, Writing-Original draft, Review & Editing. **Xi Yu:** Project administration, Supervision Funding acquisition. **Quanzhou Luo:** Methodology, Data curation, Validation. **Yujie Liu:** Methodology, Data curation, Validation. **Boyan Zhang:** Data curation, Visualization, Writing-Original draft, Review & Editing. **Haojun Gao:** Data curation, Visualization, Writing-Original draft, Review & Editing. **Yuefei Wang:** Conceptualization, Methodology, Project administration, Supervision, Review & Editing, Funding acquisition.

Acknowledgment

This research was supported by the Sichuan Comic and Animation Research Center, Key Research Institute of Social Sciences of Sichuan Province (DM2024013), the Key Laboratory of Pattern Recognition and Intelligent Information Processing of Sichuan, Chengdu University, Chengdu, China (MSSB-2025-16), and the Open Project Fund of Key Laboratory of Digital Innovation of Tianfu Culture, Sichuan Provincial Department of Culture and Tourism (TFWH-

2025-12). In addition, the computational platform for this study was supported by the National Supercomputing Center in Chengdu-Chengdu University Branch.

REFERENCES

[1] L. Deng, D. Yu, Deep Learning: Methods and Applications, Found. Trends® Signal Process. 7 (2014) 197-387. https://doi.org/10.1561/2000000039.

[2] M.R. Karim, O. Beyan, A. Zappa, I.G. Costa, D. Rebholz-Schuhmann, M. Cochez, S. Decker, Deep learning-based clustering approaches for bioinformatics, (n.d.).

[3] J.-T. Chien, Deep Bayesian Mining, Learning and Understanding, in: Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., ACM, Anchorage AK USA, 2019: pp. 3197-3198. https://doi.org/10.1145/3292500.3332267.

[4] X. Wei, Z. Zhang, H. Huang, Y. Zhou, An overview on deep clustering, Neurocomputing 590 (2024) 127761. https://doi.org/10.1016/j.neucom.2024.127761.

[5] Z. Hu, Y. Wang, H. Ning, D. Wu, F. Nie, Mutual-Taught Deep Clustering, Knowl.-Based Syst. 282 (2023) 111100. https://doi.org/10.1016/j.knosys.2023.111100.

[6] J. Wen, G. Xu, Z. Tang, W. Wang, L. Fei, Y. Xu, Graph Regularized and Feature Aware Matrix Factorization for Robust Incomplete Multi-View Clustering, IEEE Trans. Circuits Syst. Video Technol. 34 (2024) 3728-3741. https://doi.org/10.1109/TCSVT.2023.3317877.

[7] Z. Chen, Y. Li, K. Lou, L. Zhao, Incomplete Multi-View Clustering With Complete View Guidance, IEEE Signal Process. Lett. 30 (2023) 1247-1251. https://doi.org/10.1109/LSP.2023.3302234.

[8] Y. Zhang, K. Song, X. Cai, Y. Tuergong, L. Yuan, Y. Zhang, Multimodal Topic Detection in Social Networks with Graph Fusion, in: C. Xing, X. Fu, Y. Zhang, G. Zhang, C. Borjigin (Eds.), Web Inf. Syst. Appl., Springer International Publishing, Cham, 2021: pp. 28-38. https://doi.org/10.1007/978-3-030-87571-8_3.

[9] G. Du, L. Zhou, Y. Yang, K. Lü, L. Wang, Deep Multiple Auto-Encoder-Based Multi-view Clustering, Data Sci. Eng. 6 (2021) 323-338. https://doi.org/10.1007/s41019-021-00159-z.

[10] Y. Tian, D. Krishnan, P. Isola, Contrastive Multiview Coding, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Comput. Vis. - ECCV 2020, Springer International Publishing, Cham, 2020: pp. 776-794. https://doi.org/10.1007/978-3-030-58621-8_45.

[11] H. Yuan, S. Lai, X. Li, J. Dai, Y. Sun, Z. Ren, Robust Prototype Completion for Incomplete Multi-view Clustering, in: Proc. 32nd ACM Int. Conf. Multimed., ACM, Melbourne VIC Australia, 2024: pp. 10402-10411. https://doi.org/10.1145/3664647.3681397.

[12] Y. Zhao, L. Bai, Contrastive clustering with a graph consistency constraint, Pattern Recognit. 146 (2024) 110032. https://doi.org/10.1016/j.patcog.2023.110032.

[13] Y. Li, M. Yang, D. Peng, T. Li, J. Huang, X. Peng, Twin Contrastive Learning for Online Clustering, Int. J. Comput. Vis. 130 (2022) 2205-2221. https://doi.org/10.1007/s11263-022-01639-z.

[14] J.-T. Chien, K. Chen, False Negative Masking for Debiasing in Contrastive Learning, in: 2024 Int. Jt. Conf. Neural Netw. IJCNN, 2024: pp. 1-6. https://doi.org/10.1109/IJCNN60899.2024.10651188.

[15] Q. Wu, Y. Yang, Y. Jiang, Contrastive Clustering with False Negatives Exclusion and Filtering Attraction, in: 2023 18th Int. Conf. Intell. Syst. Knowl. Eng. ISKE, 2023: pp. 155-162. https://doi.org/10.1109/ISKE60036.2023.10481035.

[16] P. Su, Y. Liu, S. Li, S. Huang, J. Lv, Robust contrastive multi-view kernel clustering, in: Proc. Thirty-Third Int. Jt. Conf. Artif. Intell., Jeju, Korea, 2024. https://doi.org/10.24963/ijcai.2024/546.

[17] Y. Wang, L. Chen, Multi-exemplar based clustering for imbalanced data, in: 2014 13th Int. Conf. Control Autom. Robot. Vis. ICARCV, 2014: pp. 1068-1073. https://doi.org/10.1109/ICARCV.2014.7064454.

[18] Y. Zhang, T. Zhang, F. Ma, X. Zhang, An Improved Interval Type-2 Rough Fuzzy K-means Based on Local Imbalanced Metric, in: 2022 China Autom. Congr. CAC, 2022: pp. 473-477. https://doi.org/10.1109/CAC57257.2022.10054897.

[19] S. Liu, Q. Liao, S. Wang, X. Liu, E. Zhu, Robust and Consistent Anchor Graph Learning for Multi-View Clustering, IEEE Trans. Knowl. Data Eng. 36 (2024) 4207-4219. https://doi.org/10.1109/TKDE.2024.3364663.

[20] M.-S. Chen, J.-Q. Lin, X.-L. Li, B.-Y. Liu, C.-D. Wang, D. Huang, J.-H. Lai, Representation Learning in Multi-view Clustering: A Literature Review, Data Sci. Eng. 7 (2022) 225-241. https://doi.org/10.1007/s41019-022-00190-8.

[21] Z. Huang, J.T. Zhou, H. Zhu, C. Zhang, J. Lv, X. Peng, Deep Spectral Representation Learning From Multi-View Data, IEEE Trans. Image Process. 30 (2021) 5352-5362. https://doi.org/10.1109/TIP.2021.3083072.

[22] Q. Wang, Z. Tao, W. Xia, Q. Gao, X. Cao, L. Jiao, Adversarial Multiview Clustering Networks With Adaptive Fusion, IEEE Trans. Neural Netw. Learn. Syst. 34 (2023) 7635-7647. https://doi.org/10.1109/TNNLS.2022.3145048.

[23] S. Luo, C. Zhang, W. Zhang, X. Cao, Consistent and specific multi-view subspace clustering, in: Proc. Thirty-Second AAAI Conf. Artif. Intell. Thirtieth Innov. Appl. Artif. Intell. Conf. Eighth AAAI Symp. Educ. Adv. Artif. Intell., AAAI Press, New Orleans, Louisiana, USA, 2018.

[24] X. Cao, C. Zhang, H. Fu, S. Liu, H. Zhang, Diversity-Induced Multi-View Subspace Clustering, in: 2015: pp. 586-594.

[25] C. Zhang, H. Fu, Q. Hu, X. Cao, Y. Xie, D. Tao, D. Xu, Generalized Latent Multi-View Subspace Clustering, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2020) 86-99. https://doi.org/10.1109/TPAMI.2018.2877660.

[26] Y. Li, F. Nie, H. Huang, J. Huang, Large-Scale Multi-View Spectral Clustering via Bipartite Graph, Proc. AAAI Conf. Artif. Intell. 29 (2015). https://doi.org/10.1609/aaai.v29i1.9598.

[27] F. Nie, J. Li, X. Li, Self-weighted multiview clustering with multiple graphs, in: Proc. 26th Int. Jt. Conf. Artif. Intell., AAAI Press, Melbourne, Australia, 2017: pp. 2564-2570.

[28] X. Liu, Y. Dou, J. Yin, L. Wang, E. Zhu, Multiple Kernel *k*-Means Clustering with Matrix-Induced Regularization, Proc. AAAI Conf. Artif. Intell. 30 (2016). https://doi.org/10.1609/aaai.v30i1.10249.

[29] L. Du, P. Zhou, L. Shi, H. Wang, M. Fan, W. Wang, Y.-D. Shen, Robust multiple kernel k-means using l21-norm, 2015.

[30] M.-S. Chen, L. Huang, C.-D. Wang, D. Huang, J.-H. Lai, Relaxed multi-view clustering in latent embedding space, Inf. Fusion 68 (2021) 8-21. https://doi.org/10.1016/j.inffus.2020.10.013.

[31] M. Abavisani, V.M. Patel, Deep Multimodal Subspace Clustering Networks, IEEE J. Sel. Top. Signal Process. 12 (2018) 1601-1614. https://doi.org/10.1109/JSTSP.2018.2875385.

[32] C. Zhang, Y. Geng, Z. Han, Y. Liu, H. Fu, Q. Hu, Autoencoder in Autoencoder Networks, IEEE Trans. Neural Netw. Learn. Syst. 35 (2024) 2263-2275. https://doi.org/10.1109/TNNLS.2022.3189239.

[33] D.J. Trosten, S. Lokse, R. Jenssen, M. Kampffmeyer, Reconsidering Representation Alignment for Multi-View Clustering, in: 2021: pp. 1255-1265.

[34] Q. Gao, H. Lian, Q. Wang, G. Sun, Cross-Modal Subspace Clustering via Deep Canonical Correlation Analysis, Proc. AAAI Conf. Artif. Intell. 34 (2020) 3938-3945. https://doi.org/10.1609/aaai.v34i04.5808.

[35] J. Cheng, Q. Wang, Z. Tao, D. Xie, Q. Gao, Multi-view attribute graph convolution networks for clustering, in: Proc. Twenty-Ninth Int. Jt. Conf. Artif. Intell., Yokohama, Yokohama, Japan, 2021.

[36] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, Y. Wei, Circle Loss: A Unified Perspective of Pair Similarity Optimization, in: 2020: pp. 6398-6407.

[37] Y. Xie, B. Lin, Y. Qu, C. Li, W. Zhang, L. Ma, Y. Wen, D. Tao, Joint Deep Multi-View Learning for Image Clustering, IEEE Trans. Knowl. Data Eng. 33 (2021) 3594-3606. https://doi.org/10.1109/TKDE.2020.2973981.

[38] B. Zhang, L. Wang, False Negative Sample Detection for Graph Contrastive Learning, Tsinghua Sci. Technol. 29 (2024) 529-542. https://doi.org/10.26599/TST.2023.9010043.

[39] J.-T. Chien, K. Chen, False Negative Masking for Debiasing in Contrastive Learning, in: 2024 Int. Jt. Conf. Neural Netw. IJCNN, 2024: pp. 1-6. https://doi.org/10.1109/IJCNN60899.2024.10651188.

[40] B. Li, Y. Wang, B. Zhao, H. Gao, B. Yang, Q. Luo, X. Li, X. Xiang, Y. Liu, H. Tang, Attributed Graph Clustering with Multi-Scale Weight-Based Pairwise Coarsening and Contrastive Learning, Neurocomputing (2025) 130796.

[41] J.D. Robinson, C.-Y. Chuang, S. Sra, S. Jegelka, Contrastive Learning with Hard Negative Samples, in: 2020.

[42] R. Jiang, T. Nguyen, P. Ishwar, S. Aeron, Supervised Contrastive Learning with Hard Negative Samples, in: 2024 Int. Jt. Conf. Neural Netw. IJCNN, 2024: pp. 1-8. https://doi.org/10.1109/IJCNN60899.2024.10650863.

[43] X. Gao, M.I. Ramli, M.M. Rosli, N. Jamil, S.M.Z.S.Z. Ariffin, Revisiting self-supervised contrastive learning for imbalanced classification, Int. J. Electr. Comput. Eng. IJECE 15 (2025) 1949. https://doi.org/10.11591/ijece.v15i2.pp1949-1960.

[44] T. Tatusova, M. DiCuccio, A. Badretdin, V. Chetvernin, E.P. Nawrocki, L. Zaslavsky, A. Lomsadze, K.D. Pruitt, M. Borodovsky, J. Ostell, NCBI prokaryotic genome annotation pipeline, Nucleic Acids Res. 44 (2016) 6614-6624. https://doi.org/10.1093/nar/gkw569.

[45] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor Segmentation and Support Inference from RGBD Images, in: A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, C. Schmid (Eds.), Comput. Vis. - ECCV 2012, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012: pp. 746-760. https://doi.org/10.1007/978-3-642-33715-4_54.

[46] A.K. McCallum, K. Nigam, J. Rennie, K. Seymore, Automating the Construction of Internet Portals with Machine Learning, Inf. Retr. 3 (2000) 127-163. https://doi.org/10.1023/A:1009953814988.

[47] Y.-G. Jiang, J. Yang, C.-W. Ngo, A.G. Hauptmann, Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study, IEEE Trans. Multimed. 12 (2010) 42-53. https://doi.org/10.1109/TMM.2009.2036235.

[48] C. Beaulac, J.S. Rosenthal, Introducing a New High-Resolution Handwritten Digits Data Set with Writer Characteristics, SN Comput. Sci. 4 (2022) 66. https://doi.org/10.1007/s42979-022-01494-2.

[49] W. Xu, X. Liu, Y. Gong, Document clustering based on non-negative matrix factorization, in: Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Dev. Informaion Retr., Association for Computing Machinery, New York, NY, USA, 2003: pp. 267-273. https://doi.org/10.1145/860435.860485.

[50] A. Strehl, J. Ghosh, Cluster Ensembles --- A Knowledge Reuse Framework for Combining Multiple Partitions, J. Mach. Learn. Res. 3 (2002) 583-617.

[51] C.D. Manning, P. Raghavan, H. Schütze, Introduction to information retrieval, Cambridge university press, Cambridge, 2008.

[52] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1985) 193-218. https://doi.org/10.1007/BF01908075.

[53] D.M.W. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, (2020). https://doi.org/10.48550/ARXIV.2010.16061.

[54] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: 1967.

[55] J. Xu, Y. Ren, G. Li, L. Pan, C. Zhu, Z. Xu, Deep embedded multi-view clustering with collaborative training, Inf. Sci. 573 (2021) 279-290. https://doi.org/10.1016/j.ins.2020.12.073.

[56] Z. Lin, Z. Kang, Graph Filter-based Multi-view Attributed Graph Clustering, in: Proc. Thirtieth Int. Jt. Conf. Artif. Intell., International Joint Conferences on Artificial Intelligence Organization, Montreal, Canada, 2021: pp. 2723-2729. https://doi.org/10.24963/ijcai.2021/375.

[57] J. Xu, H. Tang, Y. Ren, L. Peng, X. Zhu, L. He, Multi-Level Feature Learning for Contrastive Multi-View Clustering, in: 2022: pp. 16051-16060.

[58] X. Yang, J. Jiaqi, S. Wang, K. Liang, Y. Liu, Y. Wen, S. Liu, S. Zhou, X. Liu, E. Zhu, DealMVC: Dual Contrastive Calibration for Multi-view Clustering, in: Proc. 31st ACM Int. Conf. Multimed., Association for Computing Machinery, New York, NY, USA, 2023: pp. 337-346. https://doi.org/10.1145/3581783.3611951.

[59] W. Yan, Y. Zhang, C. Lv, C. Tang, G. Yue, L. Liao, W. Lin, GCFAgg: Global and Cross-View Feature Aggregation for Multi-View Clustering, in: 2023: pp. 19863-19872.

[60] J. Cui, Y. Li, H. Huang, J. Wen, Dual Contrast-Driven Deep Multi-View Clustering, IEEE Trans. Image Process. 33 (2024) 4753-4764. https://doi.org/10.1109/TIP.2024.3444269.

[61] S. Wu, Y. Zheng, Y. Ren, J. He, X. Pu, S. Huang, Z. Hao, L. He, Self-weighted contrastive fusion for deep multi-view clustering, IEEE Trans. Multimed. 26 (2024) 9150-9162.

[62] W. Yan, Y. Zhang, C. Tang, W. Zhou, W. Lin, Anchor-Sharing and Cluster-Wise Contrastive Network for Multiview Representation Learning,

IEEE Trans. Neural Netw. Learn. Syst. 36 (2025) 3797-3807. https://doi.org/10.1109/TNNLS.2024.3357087.

Appendix

Main Theorem A (Directional Upper Bound: FN will not push the anchor point away from the 'true center' in the direction of divergence)

Assumptions

The representation vectors used for comparison are all ℓ_2 normalized, such that $\| z_i \| = \| z_j \| = \| u_c \| = 1$, and the similarity is defined as $s_{ij} = \cos(z_i, z_j) = z_i^\top z_j$.

A.2 (Weighted InfoNCE)

The loss for anchor i is given by

$$\ell_i = -\log \frac{\sum_{p\in P(i)} w_{ip}\, e^{\frac{s_{ip}}{\tau}}}{\sum_{r\neq i} \omega_{ir}\, e^{\frac{s_{ir}}{\tau}}}, \tag{A.1}$$

where the temperature $\tau > 0$; $P(i)$ is the set of positive indices (guided by pseudo-labels/consensus); $\omega_{ir} \geq 0$ is the denominator weight.

A.3 (FN Soft Masking and Gating)

For the sample set $\mathrm{FN}(i)$ classified as same cluster but falling into the denominator' by CSD, the denominator weight is set as follows:

$$\omega_{ir} = g\pi_{\mathrm{FN}} \quad \big(r \in \mathrm{FN}(i)\big), \qquad \omega_{ir} = 1 \quad \big(r \notin \mathrm{FN}(i)\big), \tag{A.2}$$

where $\pi_{\mathrm{FN}} \in [0,1]$ represents the soft masking coefficient, and $g \in [0,1]$ denotes the gating factor, controlled by the batch's average uncertainty, among other factors.

A.4 (Non-Degenerative)

$\boldsymbol{\sum_{r\notin \mathrm{FN}(i)} e^{s_{ir}/\tau} > 0.}$

A.5 (Directional Geometric Margin)

Let $c^{*}*$ represent the "true/target cluster" center for anchor i (which can be approximated by a stable EMA center or consensus center). Define the "unit direction away from the true center."

$$d_i^{(+)} := -\frac{\nabla_{z_i} s_{i,u_{c^*}}}{\| \nabla_{z_i} s_{i,u_{c^*}} \|} = -\frac{u_{c^*} - s_{i,u_{c^*}} z_i}{\sqrt{1 - s_{i,u_{c^*}}^2}}, \tag{A.3}$$

Where $s_{i,u_{c^*}} = z_i^\top u_{c^*}$, and denote it as $D_i := \sqrt{1 - s_{i,u_{c^*}}^2} \in (0,1]$.

Define the directional ratio for any unit vector b as:

$$A_i(b) \coloneqq \langle b, u_{c^*} \rangle - s_{i,u_{c^*}} \langle z_i, b \rangle. \tag{A.4}$$

There exist constants$\mu_+ > \mu_- \geq 0$ and $\mu_{\mathrm{FN}} \geq \mu_+$ such that :

$$\mathbb{E}_{p\in P(i)}\big[A_i(z_p)\big] \geq \mu_+, \qquad \mathbb{E}_{r\notin \mathrm{FN}(i)}[A_i(z_r)] \leq \mu_-, \qquad \mathbb{E}_{j\in \mathrm{FN}(i)}\big[A_i(z_j)\big] \geq \mu_{\mathrm{FN}}.$$

This is a mild directional separability condition: positive examples align more closely with the center, non-FN negative examples are misaligned with the center ($\leq \mu_-$), while FN (same cluster but falling into the denominator) are highly aligned with the center ($\geq \mu_{\mathrm{FN}}$). These quantities can be estimated through intra-batch or cross-batch EMA.

A.6 (FN Share)

The mass share of FN under the denominator softmax.

$$\theta_i \coloneqq \sum_{j\in \mathrm{FN}(i)} \tilde{p}_{ij} \quad \text{Within} \quad \tilde{p}_{ir} = \frac{\omega_{ir} e^{\frac{s_{ir}}{\tau}}}{\sum_{u\neq i} \omega_{iu}\, e^{\frac{s_{iu}}{\tau}}} \tag{A.5}$$

Satisfied

$$\theta_i = \frac{g\pi_{\mathrm{FN}}}{g\pi_{\mathrm{FN}} + \kappa_i}, \tag{A.6}$$

$$\kappa_i \coloneqq \frac{\sum_{r\notin \mathrm{FN}(i)} e^{\frac{s_{ir}}{\tau}}}{\sum_{j\in \mathrm{FN}(i)} e^{\frac{s_{ij}}{\tau}}} \in (0, +\infty], \tag{A.7}$$

It automatically holds true by (A.2) and (A.4).

Theorem A (Directional Upper Bound: A single update in the direction "opposite to the true center" does not yield a positive/negative margin)

Let a parameter update be defined as $z_i^{\mathrm{new}} = z_i - \eta \nabla_{z_i} \ell_i$ followed by re-normalization (projected onto the unit sphere; this projection has no effect on the first-order direction projection). Under assumptions A.1-A.6, the closed-form expression for the directional projection and its upper bound is given by:

$$\left\langle z_i^{\mathrm{new}} - z_i, d_i^{(+)} \right\rangle = -\frac{\eta}{\tau D_i} \left[\underbrace{\sum_{p\in P(i)} q_{ip} A_i(z_p)}_{\substack{\text{Numerator (positive examples)} \\ \text{attracts – 'good terms'}}} \underbrace{\sum_{r\neq i} \tilde{p}_{ir} A_i(z_r)}_{\substack{\text{Denominator (including FN)} \\ \text{repels – 'bad terms'}}} \right], \tag{A.8}$$

Where $q_{ip} = \frac{w_{ip} e^{s_{ip}/\tau}}{\sum_{u\in P(i)} w_{iu} e^{s_{iu}/\tau}}$. Furthermore, define:

$$\Xi_i \coloneqq \mu_- + \theta_i (\mu_{\mathrm{FN}} - \mu_-) - \mu_+, \tag{A.9}$$

Then,we have

$$\left\langle z_i^{\mathrm{new}} - z_i, d_i^{(+)} \right\rangle \leq -\frac{\eta}{\tau D_i} \Xi_i. \tag{A.10}$$

In particular, if

$$\theta_i \leq \frac{\mu_+ - \mu_-}{\mu_{\mathrm{FN}} - \mu_-} \Leftrightarrow g\pi_{\mathrm{FN}} \leq \frac{\mu_+ - \mu_-}{\mu_{\mathrm{FN}} - \mu_+}\kappa_i, \tag{A.11}$$

Then $\Xi_i \leq 0$, thus:

$$\left\langle z_i^{\mathrm{new}} - z_i, d_i^{(+)} \right\rangle \leq 0,$$

That is the update will not proceed in the direction "opposite to the true center"; if it is strictly smaller, there exists a negative margin (on the side toward the true center). The directional projection (A.10) simultaneously incorporates the contributions from both the denominator (including FN) and the numerator (positive examples). As long as the softmax share θ_i of FN, controlled by π_{FN}, g, and κ_i, does not exceed the threshold in (A.11), the overall update will not push the anchor point away from its "true center." The positive example term in the numerator "counteracts/suppresses" the negative impact of FN in the direction.

Proof

(1) Directional Inner Product Identity

From A.1 $\Rightarrow \nabla_{z_i} s_{ib} = b - s_{ib} z_i$, and

$$d_i^{(+)} = -\frac{\nabla_{z_i} s_{i,u_c^*}}{\parallel \nabla_{z_i} s_{i,u_c^*} \parallel} = -\frac{u_c^* - s_{i,u_c^*} z_i}{D_i}, \quad D_i = \sqrt{1 - s_{i,u_c^*}^2}.$$

for any unit vector b:

$$\left\langle \nabla_{z_i} s_{ib}, d_i^{(+)} \right\rangle = -\frac{\langle b - s_{ib} z_i, u_c^* - s_{i,u_c^*} z_i \rangle}{D_i} = -\frac{\langle b, u_c^* \rangle - s_{i,u_c^*} \langle z_i, b \rangle}{D_i} = -\frac{A_i(b)}{D_i}. \tag{$A.12$}$$

(2) Gradient Decomposition, Explicit "Numerator − Denominator"

Direct differentiation from (A.1) (standard InfoNCE calculation):

$$\nabla_{z_i} \ell_i = \frac{1}{\tau}\left(\sum_{r \neq i} \tilde{p}_{ir} \, \nabla s_{ir} - \sum_{p \in P(i)} q_{ip} \, \nabla s_{ip} \right). \tag{$A.13$}$$

The direction projection of a single update

$$\left\langle z_i^{\mathrm{new}} - z_i, d_i^{(+)} \right\rangle = -\eta \left\langle \nabla_{z_i} \ell_i, d_i^{(+)} \right\rangle. \tag{$A.14$}$$

Substituting (A.13) and (A.12) yields the closed-form expression (A.8):

$$-\eta \cdot \frac{1}{\tau}\left(\sum_{r \neq i} \tilde{p}_{ir} \left(-\frac{A_i(z_r)}{D_i} \right) - \sum_{p \in P(i)} q_{ip} \left(-\frac{A_i(z_p)}{D_i} \right) \right) = -\frac{\eta}{\tau D_i}\left[\sum_p q_{ip} \, A_i(z_p) - \sum_r \tilde{p}_{ir} \, A_i(z_r) \right].$$

(3) Directional Expectation and Upper Bound

The denominator and the expression are split into $FN(i)$ / non-FN, and the expectation lower/upper bounds are taken (Jensen):

$$\sum_r \tilde{p}_{ir}\, A_i(z_r) = \theta_i \cdot \mathbb{E}_{j\in\mathrm{FN}}\big[A_i(z_j)\big] + (1-\theta_i)\cdot \mathbb{E}_{r\notin\mathrm{FN}}[A_i(z_r)] \ge \theta_i\mu_{\mathrm{FN}} + (1-\theta_i)\mu_-$$

The numerator term is the weighted average:

$$\sum_p q_{ip}\, A_i\big(z_p\big) \ge \mu_+$$

Substituting back into (A.8) results in (A.10):

$$\left\langle z_i^{\mathrm{new}} - z_i, d_i^{(+)}\right\rangle \le -\frac{\eta}{\tau D_i}(\mu_+ - [\theta_i\mu_{\mathrm{FN}} + (1-\theta_i)\mu_-]) = -\frac{\eta}{\tau D_i}\Xi_i$$

(4) Threshold Condition Equivalence

$$\Xi_i \le 0 \Leftrightarrow \theta_i \le \frac{\mu_+ - \mu_-}{\mu_{\mathrm{FN}} - \mu_-}$$

From (A.6) 的 $\theta_i = \frac{g\pi_{\mathrm{FN}}}{g\pi_{\mathrm{FN}}+\kappa_i}$ we obtaion:

$$\frac{g\pi_{\mathrm{FN}}}{g\pi_{\mathrm{FN}}+\kappa_i} \le \frac{\mu_+ - \mu_-}{\mu_{\mathrm{FN}} - \mu_-} \Leftrightarrow g\pi_{\mathrm{FN}} \le \frac{\mu_+ - \mu_-}{\mu_{\mathrm{FN}} - \mu_+}\kappa_i$$

Thus (A.11); therefore, the projection ≤ 0.Q.E.D

Main Theorem B (Stable Center + Preservation of Separation + Distribution Alignment)

Assumptions

B.1 (Minority Cluster and Geometric Distance within the Batch)

Let there exist a set of minority clusters $M \subset \{1, \dots, C\}$ within the current batch. Denote the global minimum center distance as

$$\gamma \triangleq \min_{c\neq c'} \| u_c - u_{c'} \| > 0. \qquad \text{(B.1)}$$

B.2 (Pseudo-Sample Generation and Covariance Amplification)

For each minority cluster $c \in M$, generate P pseudo-samples in the consensus space:

$$\psi_{c,p} \sim \mathcal{N}(\mu_c, \Sigma_c{}'), \qquad \Sigma_c' = \frac{P}{n_c}\Sigma_c + \varepsilon I_D, \quad \varepsilon > 0, \qquad \text{(B.2)}$$

where μ_c, Σ_c represent the empirical mean and covariance of the real samples in the batch (Equations (25) and (26)), n_c is the number of real samples in the batch, and I_D is the $D \times D$ identity matrix.

B.3 (MMD Alignment and Kernel)

Using the RBF kernel $K_\sigma(x, y) = \exp\big(-\| x - y \|^2/(2\sigma^2)\big)$ with $\sigma > 0$, define the batch-wise MMD as

$$D_{\mathrm{MMD}}^{(c)} = \tfrac{1}{n_c(n_c-1)}\textstyle\sum_{i\neq j} K_\sigma\left(z_i, z_j\right) + \tfrac{1}{P(P-1)}\textstyle\sum_{p\neq q} K_\sigma\left(\psi_{c,p}, \psi_{c,q}\right) - \tfrac{2}{n_c P}\textstyle\sum_{i,p} K_\sigma\left(z_i, \psi_{c,p}\right). \qquad \text{(B.3)}$$

The RBF kernel is characteristic.

B.4 (Repulsion Regularization and Boundary Constraints)

The Gaussian repulsion loss is defined as

$$L_{\text{rep}} = \sum_{c\neq c'}\sum_{p,q}\exp\left(-\frac{\left\|\psi_{c,p}-\psi_{c',q}\right\|^2}{2\sigma_r^2}\right),\quad \sigma_r>0, \tag{B.4}$$

The boundary constraint is given by

$$L_b = \sum_{c\in M}\sum_{p=1}^{P}\left[\max\left(\|\psi_{c,p}-u_c\|-R_{\max},0\right)+\max\left(R'_{\min}-\min_{j\neq c}\|\psi_{c,p}-u_j\|,0\right)\right]. \tag{B.5}$$

Set controllable thresholds $R_{\max}, R'_{\min}$ such that

$$R_{\max} < \frac{\gamma}{2},\quad R'_{\min} > \frac{\gamma}{2}.$$

B.5 (ISM Total Regularization)

The ISM regularization is defined as

$$L_{\text{imb}} = \sum_{c\in M}\lambda_{\text{mmd}}\,D_{\text{MMD}}^{(c)} + \lambda_{\text{rep}}\,L_{\text{rep}} + \lambda_b L_b,\quad \lambda_{\text{mmd}},\lambda_{\text{rep}},\lambda_b \geq 0. \tag{B.6}$$

The above assumptions only involve batch-wise statistics, kernels, and geometric thresholds, with no strong distributional assumptions.

Theorem B (Stable Center + Classification Preservation + Distribution Alignment)

Under B.1 to B.5, for any minority cluster $c\in M$:

(B1) Variance Reduction in Center Estimation (computable, with hyperparameters)

Re-estimate the center after incorporating the pseudo-samples.

$$\hat{\mu}_c = \frac{1}{n_c+P}\left(\sum_{i=1}^{n_c} z_i + \sum_{p=1}^{P}\psi_{c,p}\right), \tag{B.7}$$

Thus,

$$\mathrm{E}[\hat{\mu}_c] = \mu_c,\qquad \mathrm{Var}(\hat{\mu}_c) = \frac{n_c\Sigma_c + P\Sigma'_c}{(n_c+P)^2} = \underbrace{\frac{n_c+\frac{P^2}{n_c}}{(n_c+P)^2}}_{\phi(P;n_c)}\Sigma_c + \frac{P}{(n_c+P)^2}\varepsilon I_D. \tag{B.8}$$

$$\phi(P;n_c) = \frac{n_c+\frac{P^2}{n_c}}{(n_c+P)^2}. \tag{B.9}$$

When ε is sufficiently small and $P=n_c$,

$$\mathrm{Var}(\hat{\mu}_c) \preccurlyeq \frac{1}{2n_c}\Sigma_c + \frac{\varepsilon}{4n_c}I_D, \tag{B.10}$$

Compared to the variance of the empirical mean using only real samples, Σ_c/n_c, this achieves at least a 2-fold variance reduction (with ε small, it approximately reaches a 2-fold reduction).

(B2) Preservation of Separability (Safety Radius Margin)

If the optimization achieves $L_b = 0$ (i.e., all pseudo-samples satisfy the inequality in (B.4)) and the threshold satisfies (B.5), then for each pseudo-sample $\psi_{c,p}$:

$$\underset{j}{\operatorname{argmin}} \parallel \psi_{c,p} - u_j \parallel = c, \tag{B.11}$$

And for any $j \neq c$,

$$\parallel \psi_{c,p} - u_j \parallel - \parallel \psi_{c,p} - u_c \parallel \geq \gamma - 2R_{\max} > 0. \tag{B.12}$$

This means that the pseudo-samples will not cross clusters and will retain at least $\gamma - 2R_{\max}$ of the geometric margin.

(B3) MMD Alignment and Distribution Consistency (Characteristic Kernel)

If the optimization ensures that

$\sum_{c\in M} D_{\mathrm{MMD}}^{(c)} \leq \delta$ $(\delta \geq 0)$, then for each $c \in M$, the kernel mean embedding distance between the pseudo-sample empirical distribution $\widehat{\Psi}_c$ and the true empirical distribution $\widehat{D}_c$ satisfies

$\parallel \mu_K(\widehat{\Psi}_c) - \mu_K(\widehat{D}_c) \parallel_{\mathcal{H}} \leq \sqrt{\delta_c}$, and as $\delta_c \to 0$ (e.g., when $\lambda_{\mathrm{mmd}}$is sufficiently large or optimization is sufficient),

$$\widehat{\Psi}_c \Rightarrow \widehat{D}_c \quad \text{(Weak Convergence)}.$$

This ensures that the pseudo-samples will not systematically deviate from the support of the true cluster distribution (in the same kernel sense).

Proof of Theorem B

Proof of (B1): Variance Reduction

From (B.7) and the linear expectation, we obtain $\mathbb{E}[\hat{\mu}_c] = \mu_c$. The covariance is given by:

$$\mathrm{Var}(\hat{\mu}_c) = \frac{1}{(n_c + P)^2}\left(\sum_{i=1}^{n_c} \mathrm{Var}\,(z_i) + \sum_{p=1}^{P} \mathrm{Var}\left(\psi_{c,p}\right)\right) = \frac{n_c\Sigma_c + P\Sigma_c'}{(n_c + P)^2}.$$

Substituting into (B.1) yields (B.8). Let $\phi(P; n_c) = (n_c + P^2/n_c)/(n_c + P)^2$. To find the extreme value of $P \geq 0$ (or set $p = P/n_c$), we verify that ϕ achieves its minimum when $P = n_c$:

$$\phi(n_c; n_c) = \frac{n_c + n_c}{(2n_c)^2} = \frac{1}{2n_c}.$$

Thus, we have (B.10). When ε is sufficiently small (in practice, ε is a numerical stability term), the second term approximately becomes 0, indicating that at least a 2-fold reduction holds. The proof is complete.

Additionally, as $P \to \infty$, $\phi(P; n_c) \to 1/n_c$, meaning that an excess of pseudo-samples does not continue to provide benefits. This also explains why, in practice, P is most appropriately chosen to be of the same order as n_c.

Proof of (B2): Preserving Separability

If $L_b = 0$, then for each $\psi_{c,p}$, we have:

$$\parallel \psi_{c,p} - u_c \parallel \leq R_{\max}, \quad \min_{j\neq c} \parallel \psi_{c,p} - u_j \parallel \geq R_{\min}{}'$$

From (B.5) and the triangle inequality, for any $j \neq c$:

$$\| \psi_{c,p} - u_j \| \geq \| \psi_{c,p} - u_c \| - \| \psi_{c,p} - u_j \| \geq \gamma - R_{\max} > \frac{\gamma}{2}.$$

Also, $\| \psi_{c,p} - u_c \| \leq R_{\max} < \gamma/2$. Therefore,

$$\| \psi_{c,p} - u_j \| - \| \psi_{c,p} - u_c \| \geq \gamma - 2R_{\max} > 0,$$

which gives (B.11) and (B.12). The proof is complete.

Role of the Exclusion Term:$\lambda_{rep} > 0$, in conjunction with σ_r, ensures that when pseudo-samples from different minority clusters become too close, an exponential penalty is applied to prevent them from crowding near the safe boundary and forming local clusters. Although the conclusion of (B.2) sufficiently ensures that pseudo-samples do not cross clusters when $L_b = 0$, λ_{rep} numerically stabilizes the optimization, preventing pseudo-samples from "clumping together."

Proof of (B3): MMD Alignment and Distribution Consistency

The Gaussian kernel is characteristic, so for any two probability measures P and Q, we have

$$\mathrm{MMD}_{\sigma}(\mathbb{P}, \mathbb{Q}) = 0 \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

The within-batch empirical quantity $D_{\mathrm{MMD}}^{(c)}$ is an unbiased estimate of the empirical MMD (ignoring the $O(1/P)$ correction). Therefore, when optimization ensures $D_{\mathrm{MMD}}^{(c)} \leq \delta_c$, the difference in the kernel mean embedding norm is given by $\| \mu_K(\widehat{\Psi}_c) - \mu_K(\widehat{D}_c) \|_{\mathcal{H}} \leq \sqrt{\delta_c}$. Taking $\delta_c \to 0$ (for example, by setting λ_{mmd} sufficiently large or P sufficiently large and optimization sufficiently thorough), by the characteristic property, we obtain $\widehat{\Psi}_c \Rightarrow \widehat{D}_c$. The proof is complete.

Main Theorem C (Under Two Time Scales, Spherical Gradient Expands the Angular Separation of HN in a Single Step)

Assumptions

C.1 (Unit Sphere and Cosine)

All vectors involved in the comparison lie on the unit sphere $\mathbb{S}^{d-1}$: $\| z_i \| = \| z_j \| = \| u_c \| = 1$. The similarity:

$$s^+ := \langle z_i, u^+ \rangle, \qquad s^- := \langle z_i, u^- \rangle.$$

C.2 (Two Time Scales: The Center is Updated Once Every T Steps)

During the inner iterations $k = 0, 1, \dots, T-1$, the centers (u^+, u^-) are regarded as constants, with only z_i being updated. In the outer loop, the centers are then collectively updated every T steps.

C.3 (HN and Margin Activation)

Sample i is identified as a HN by CSD and is currently violating the margin:

$$t := s^+ - s^- + m > 0, \qquad m > 0. \tag{C.1}$$

C.4 (Spherical (Riemannian) Gradient Descent)

A single update is performed using the spherical gradient:

$$z_i^{\mathrm{new}} = \mathrm{Exp}_{z_i}\big(-\eta \,\mathrm{grad}\, L_{\mathrm{HN}}(z_i)\big), \quad \mathrm{grad}\, f(z) = (I - zz^{\top})\nabla f(z), \tag{C.2}$$

where

$$L_{\mathrm{HN}} = g\big(\lambda_{\mathrm{push}} L_{\mathrm{push}} + \lambda_{\mathrm{pull}} L_{\mathrm{pull}}\big), \quad L_{\mathrm{push}} = \max\{0, t\}, \quad L_{\mathrm{pull}} = 1 - s^{-}, \tag{C.3}$$

$g \in [0,1], \lambda_{\mathrm{push}}, \lambda_{\mathrm{pull}} > 0$, learning rate $\eta > 0$.

C.5 (Geometric Separation of Centers)

Separation between the two centers:

$$\rho := \| u^{+} - u^{-} \| = \sqrt{2 - 2\langle u^{+}, u^{-}\rangle} \geq \rho_{\min} > 0. \tag{C.4}$$

C.6 (Riemannian Smoothness)

The function $\Delta(z) := s^{-}(z) - s^{+}(z)$ has an L_{Δ} -Lipschitz Riemannian gradient on $\mathbb{S}^{d-1}$: $\| \mathrm{grad}\Delta(z) - \mathrm{grad}\Delta(z') \| \leq L_{\Delta} \mathrm{dist}(z, z')$.

C.7 (Correctness and "HN Closer to the Wrong Cluster" Holds with Probability)

There exist$\boldsymbol{\delta_{\mathrm{sel}}}, \boldsymbol{\delta_{\mathrm{gap}}} \in [\mathbf{0}, \mathbf{1})$ such that:

a. Event $\mathcal{E}_{\mathrm{sel}}: u^{-}$ is the "correct negative cluster center" (i.e., a true center outside the ground-truth class yet geometrically the closest among them), satisfying

$\mathbf{Pr}(\boldsymbol{\mathcal{E}_{\mathrm{sel}}}) \geq \mathbf{1} - \boldsymbol{\delta_{\mathrm{sel}}}$;

b. Event $\mathcal{E}_{\mathrm{gap}}: s^{+} \geq s^{-} \geq 0$ satisfies

$\mathbf{Pr}\big(\boldsymbol{\mathcal{E}_{\mathrm{gap}}}\big) \geq \mathbf{1} - \boldsymbol{\delta_{\mathrm{gap}}}$.

Theorem C (One-Step Angular Separation Growth: Deterministic Lower Bound from Push and Non-Negative Contribution from Pull)

Let the angular separation be denoted as $\Delta := s^{-} - s^{+}$. Under assumptions C.1-C.6 , and when the event $\mathcal{E} := \mathcal{E}_{\mathrm{sel}} \cap \mathcal{E}_{\mathrm{gap}} \cap \{t > 0\}$ occurs (with probability at least $\geq 1 - \delta$, *where* $\delta := \delta_{\mathrm{sel}} + \delta_{\mathrm{gap}}$) , the one-step spherical update satisfies:

$$\Delta^{\mathrm{new}} - \Delta \geq \underbrace{\eta g \lambda_{\mathrm{push}}(\rho - m)^2}_{\substack{\text{push:} \\ \text{Deterministic Positive Gain}}} + \underbrace{\eta g \lambda_{\mathrm{pull}} \Gamma_{\mathrm{pull}}}_{\substack{\text{pull:} \\ \text{Non-Negative Contribution}}} \underbrace{- \frac{L_{\Delta}}{2} \eta^2 g^2 \big[\lambda_{\mathrm{push}}(\rho + m) + \lambda_{\mathrm{pull}}\big]^2}_{\substack{\text{Second−Order Term} \\ \text{(Smoothness Penalty)}}} \tag{C.5}$$

where

$$\Gamma_{\mathrm{pull}} \geq \left(\sqrt{1 - (s^{-})^2} - \sqrt{1 - (s^{+})^2}\right)\sqrt{1 - (s^{-})^2} \overset{\mathcal{E}_{\mathrm{gap}}}{\geq} 0. \tag{C.6}$$

In particular, when the learning rate satisfies

$$0 < \eta \leq \frac{2\lambda_{\mathrm{push}}(\rho - m)^2}{L_{\Delta} g \big[\lambda_{\mathrm{push}}(\rho + m) + \lambda_{\mathrm{pull}}\big]^2} \tag{C.7}$$

then, on event $\mathcal{E}$, it holds that $\Delta^{\text{new}} > \Delta$ (the angular separation expands monotonically). Consequently, with probability at least $1-\delta$, a single step pushes the HN sample away from the boundary toward the incorrect cluster direction, thereby alleviating overlap.

Proof

(I) Spherical Gradient and Basic Derivatives

On $\mathbb{S}^{d-1}$, $s^{\pm}(z)=\langle z,u^{\pm}\rangle$, Its Euclidean gradient is $\nabla s^{\pm}=u^{\pm}-s^{\pm}z$, which is orthogonal to z (and thus already lies in the tangent space). Therefore, the Riemannian gradient is given by grad $s^{\pm}=\nabla s^{\pm}$.

Thus

$$\text{grad}\,\Delta=\text{grad}\,s^{-}-\text{grad}\,s^{+}=(u^{-}-s^{-}z)-(u^{+}-s^{+}z)=(u^{-}-u^{+})-\Delta z. \qquad (\text{C}.8)$$

Norm Bound: $\parallel \text{grad}\,s^{\pm}\parallel=\sqrt{1-(s^{\pm})^2}\le 1.$

(II) Riemannian Gradient of the HN Loss

In the active region $t>0$,

$$\text{grad}\,L_{\text{HN}}=g\big(\lambda_{\text{push}}(\text{grad}\,s^{+}-\text{grad}\,s^{-})-\lambda_{\text{pull}}\text{grad}\,s^{-}\big). \qquad (\text{C}.9)$$

(III) First-Order Directional Gain: Push and Pull

Directional Derivative Along $-\text{grad}\,L_{\text{HN}}$

$$-\langle\text{grad}\,\Delta,\text{grad}\,L_{\text{HN}}\rangle=g\lambda_{\text{push}}\underbrace{\langle\text{grad}\,s^{+}-\text{grad}\,s^{-},\text{grad}\,\Delta\rangle}_{\text{(a)}}+g\lambda_{\text{pull}}\underbrace{\langle\text{grad}\,s^{-},\text{grad}(\Delta)\rangle}_{\text{(b)}}. \qquad (\text{C}.10)$$

For (a), since grad $\Delta=$ grad $s^{-}-$ grad s^{+} we have

$$\langle\text{grad}\,s^{+}-\text{grad}\,s^{-},\text{grad}\,\Delta\rangle=-\parallel\text{grad}\,s^{+}-\text{grad}\,s^{-}\parallel^2.$$

Furthermore, by

$$\parallel\text{grad}\,s^{+}-\text{grad}\,s^{-}\parallel=\parallel(u^{+}-u^{-})+(s^{-}-s^{+})z\parallel\ge\parallel u^{+}-u^{-}\parallel-|s^{-}-s^{+}|\ge\rho-m,$$

We have

$$(a)\ge(\rho-m)^2. \qquad (\text{C}.11)$$

For(b),

$$\langle\text{grad}s^{-},\text{grad}\Delta\rangle=\parallel\text{grad}s^{-}\parallel^2-\langle\text{grad}s^{+},\text{grad}s^{-}\rangle\ge\parallel\text{grad}s^{-}\parallel(\parallel\text{grad}s^{-}\parallel-\parallel\text{grad}s^{+}\parallel),$$

From $\parallel\text{grad}s^{\pm}\parallel=\sqrt{1-(s^{\pm})^2}$ and the event $\mathcal{E}_{\text{gap}}:s^{+}\ge s^{-}\ge 0$, it follows that

$$\parallel\text{grad}s^{-}\parallel-\parallel\text{grad}s^{+}\parallel=\sqrt{1-(s^{-})^2}-\sqrt{1-(s^{+})^2}\ge 0,$$

Thus

$$(b)\ge\Gamma_{\text{pull}}:=\left(\sqrt{1-(s^{-})^2}-\sqrt{1-(s^{+})^2}\right)\sqrt{1-(s^{-})^2}\ge 0,$$

Namely, Eq. (V.6). Substituting (C.11) and (C.6) back into (C.10) yields a first-order gain lower bound:

$$-\langle\text{grad}\Delta,\text{grad}L_{\text{HN}}\rangle\ge g\lambda_{\text{push}}(\rho-m)^2+g\lambda_{\text{pull}}\Gamma_{\text{pull}}. \qquad (\text{C}.12)$$

(IV) Riemannian Descent Lemma (Including the Second-Order Term)

For a Riemannian-smooth function, a single step along $-\eta$ grad L_{HN} via the exponential map yields

$$\Delta^{\mathrm{new}} \geq \Delta + \eta \cdot (-\langle \mathrm{grad}\Delta, \mathrm{grad}L_{\mathrm{HN}} \rangle) - \frac{L_\Delta}{2}\eta^2 \parallel \mathrm{grad}L_{\mathrm{HN}} \parallel^2. \tag{C.13}$$

Furthermore, by

$$\parallel \mathrm{grad}L_{\mathrm{HN}} \parallel \leq g\big(\lambda_{\mathrm{push}} \parallel \mathrm{grad}s^+ - \mathrm{grad}s^- \parallel + \lambda_{\mathrm{pull}} \parallel \mathrm{grad}s^- \parallel\big) \leq g\big(\lambda_{\mathrm{push}}(\rho + m) + \lambda_{\mathrm{pull}}\big) \tag{C.14}$$

Substituting (C.12) and (C.14) into (C.13) yields (C.5). Requiring the right-hand side to be > 0 and solving for an upper bound on η gives (C.7). Q.E.D.