Views Attention Fusion of Granular-ball Fuzzy Representations Split for Improved Multi-View Clustering

Shuaiyu Liu[1], Song Wu[1], Jie Xu[1,2], Yazhou Ren[1,3,*], Yang Yang[1], Xiaorong Pu[1,3], Guoyin Wang[4]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China
[2]Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore
[3]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen, China
[4]Key Laboratory of Cyberspace Big Data Intelligent Security, Ministry of Education, China

Abstract

Multi-View Clustering (MVC) is a pivotal multi-view learning paradigm widely adopted across various fields. Despite recent advances, existing methods primarily focus on enhancing the performance of fused multi-view representation, often neglecting the issue of Representation Degradation (RD) arising from discrepancies in the intrinsic quality of different views. To address the limitations, we propose a novel Granular-ball Fuzzy Split and Attention Fusion (GFSAF) learning, which leverages the nature of granular-ball to extract mutual and complementary representation separately. Meanwhile, the proposed method introduces an attention variant for fused representations to mitigate the RD issue. GFSAF mainly consists of two training stages: Split-Extract Stage and Views-Fusion Stage. Specifically, we design a novel Granular-ball Fuzzy Contrastive Learning to extract mutual representation, and introduce Noise Stripping Loss to reduce the influence of noise for complementary representation. Then, a novel multi-head Cross Views Attention is proposed to employ attention mechanism from multi-view perspectives for comprehensive fused representations. Experimental results on eight databases demonstrate that our GFSAF achieves superior performance compared to several state-of-the-art MVC methods.

Code — https://github.com/Lsy235/GFSAF

Introduction

Nowadays, in the era of highly intensive informationization, events or objects of the same type are frequently captured through multiple heterogeneous devices (Zhang et al. 2020; Sun et al. 2023; Li et al. 2023), resulting in the inherently multi-view nature of data. With the advancement of information technologies, the diversity and complexity of such multi-view data continue to increase. Consequently, effectively extracting comprehensive and valuable information from multi-view data has emerged as a core challenge for Multi-View Learning (MVL).

The essential difference between multi-view data and single-view data is that multi-view data exhibits both homogeneity and heterogeneity across views (Xu et al. 2025). Specifically, views contain homogeneous mutual information while retaining heterogeneous complementary information unique to each view. MVL focuses on effectively balancing and integrating these two types of information to achieve a more comprehensive representation. This fundamental difference in the nature of data leads to different learning paradigms. Single view approaches tend to learn cross-sample generalized representations based on mutual information. In contrast, multi-view approaches combine multiple views of data and aim to learn more comprehensive representations by effectively integrating mutual and complementary information.

Although existing methods have shown extraordinary performance, a critical issue that remains largely unaddressed is that the performance of downstream tasks is often dominated by a single superior view, which can even outperform the fusion of all views (Xu et al. 2023). For specific explanation, we consider a representative MVL task, i.e., Multi-View Clustering (MVC) (Zhang et al. 2025b; Ren et al. 2024b). The clustering performance obtained from one high-quality view often surpasses that of the multi-view fusion, which is referred to as Representations Degradation (RD). Although some methods (Xu et al. 2023; Wu et al. 2024; Su et al. 2025) consider to address the RD issue, they are still limited by the complexity of real-world multi-view data and the RD issue need to be further investigated. Essentially, the current learning paradigm of the methods fails to align with the concept of MVL, which is to enable the model to learn more comprehensive and robust understanding of the samples by utilizing diverse perspectives from multiple views.

RD phenomenon primarily arises from the model's tendency to over-rely on mutual information while insufficiently capturing complementary information. During training, the representations of each view increasingly encode mutual information, whereas complementary information diminishes. This process results in elevated similarity between the representations of high-quality and low-quality views. While such learning may enhance the performance of individual views, it ultimately leads to mediocre fusion results. Consequently, the high-quality view often outperforms the fused representation, particularly when other views are of lower quality. To establish a genuinely effective MVL paradigm, it is crucial to address two longstanding challenges: (i) the inevitable quality discrepancy among differ-

*Corresponding Author (yazhou.ren@uestc.edu.cn).

ent views, and (ii) the effective extraction and fusion of both mutual and complementary information across views.

The first challenge is commonly introduced during data collection due to factors (Ren et al. 2024a; Chen et al. 2024) such as device heterogeneity, acquisition conditions. Although the low-quality view naturally lacks sufficient information, it also contains complementary information that is crucial for MVL. We propose an innovative paradigm for complementing the information of high-quality views with the complementary information of low-quality views to cleverly deal with quality discrepancy. The second challenge lies in extraction and fusion. Mutual information pursues more robust representation, which emphasizes the common semantics captured across views. As in multi-view images of the trucks, mutual information typically denotes the outline or structural shape, whereas complementary information includes unique attributes such as colors or decorations.

For extraction, we adopt granular-ball computing (Xia et al. 2019) and introduce a novel Granular-ball Fuzzy Contrastive Learning (GFCL) for improving MVL. As in the Split-Extract Stage shown in Fig. 1, traditional method operates at the sample level, which leads to an overemphasis on the special information of each sample. Alternatively, using single-size clusters as a basis for classification is simple but tends to suboptimal solutions (Su et al. 2025). To this end, GFCL is proposed to learn representations of mutual information with stronger robustness across views at multiple granularities level. For complementary information, typically hidden in fine-grained local patterns, we employ deeper network and $\mathcal{L}_{NS}$ to split-extract representations. In the fusion, we innovatively design a Cross View Attention (CVA) module, a multi-view attention variant inspired by multi-head attention, that effectively fuses both mutual and complementary representations. Unlike usual fusion methods such as concatenation or weighted averaging (Yang et al. 2023; Luo et al. 2024; Su et al. 2025), our studies reveal that the relationship between mutual and complementary information is non-linear and highly entangled, which causes traditional methods to fall into the RD challenge.

Moreover, we propose a novel Granular-ball Fuzzy Split and Attention Fusion (GFSAF) method based on a two-stage learning paradigm, consisting of a Split-Extract Stage (SES) and a Views-Fusion Stage (VFS). Unlike traditional two-stage methods, in SES, we introduce and integrate granular-ball and multi-view contrastive learning. Via a customized pre-training structure, our model aims not only to minimize reconstruction differences, but more importantly, to explicitly split and extract mutual and complementary information across views as the primary objective. In VFS, we design a novel variant of the attention mechanism customized for MVL. Motivated by the complex interdependence between mutual and complementary information, the design of CVA enables to fuse heterogeneous and homogeneous information more effectively. The proposed method significantly alleviates the RD problem observed in existing methods.

Our main contributions can be summarized as follows.

- A novel two-stage learning paradigm for MVL is proposed, which effectively deals with the two challenges. In GFSAF, we innovatively divide MVL into the two-stage split and fusion task. In SES, one fuzzy contrastive learning in the level of granular-ball is designed to improve the ability of model, which separates and extracts the two types of information effectively.
- We propose a novel attention mechanism variant for MVL, which explicitly alleviates the problem of RD. Based on complex and non-linear interdependencies between mutual and complementary information, the proposed CVA utilizes a cross-view attention mechanism to fuse cross-view representations.
- Our method is evaluated on multiple databases compared with a variety of state-of-the-art methods in recent years, effectively mitigates RD and achieves superior results.

Related Work

Multi-View Learning

Multi-View Learning, shortly named MVL, has recently gained significant attention and has been widely applied in clustering (Jiang et al. 2025; He et al. 2025; Ren et al. 2025), recognition (Lu et al. 2025; Zhang et al. 2025a), and so on. However, how to fully and effectively utilize the information containing in multi-view data remains a problem, which is closely related to the challenges mentioned above. Su et al. (2025) proposed a contrastive clustering method and selectively deleted some views to solve quality differences. Yang et al. (2023) proposed a novel dual contrastive calibration network for optimizing cross-view learning. Sun et al. (2024) proposed RMCNC to alleviate the influence of misaligned pairs from multi-view data. Although the above methods effectively improve the information extraction ability from different perspectives, the RD problem arises, which deviates from the concept of MVL. Xu et al. (2023) proposed a novel self-weighted multi-view contrastive learning with reconstruction regularization to alleviate the RD problem.

Our method proposes a novel MVL two-stage method. We introduce multi-granularity balls to effectively alleviate the influence of quality differences and improve the effect of representations split extraction. In addition, we design a novel attention variant to fuse information representations, which effectively alleviates the RD and has been proven to achieve superior performance via extensive experiments.

Granular-ball Computing

Inspired by the "large scale first" cognitive mechanism (Chen 1982) and multi-granularity cognitive computation, Xia et al. (2019) proposed one efficient and robust method named granular-ball computing. While preserving the quality of the original database, it can greatly reduce the amount of data, which can be used to effectively mine the distribution structure between arbitrary data of the same category for more generalized models. In recent years, granular-ball has made significant strides in many fields, such as clustering (Xie et al. 2024), classification (Li et al. 2025; Huang et al. 2025), graph learning (Xia et al. 2025b), generation (Hu et al. 2025; Xia et al. 2025a), etc. Quadir and Tanveer (2024) proposed granular ball twin support vector machine to deal with challenges in TSVM field.

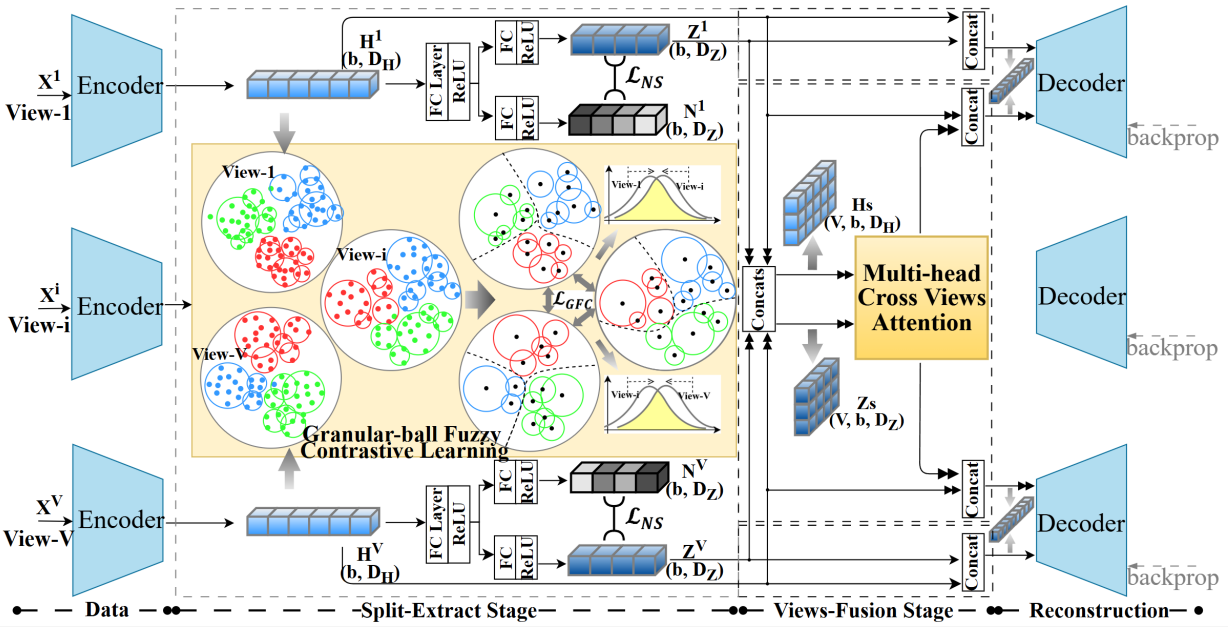


Figure 1: Overview of our proposed GFSAF. Multi-view encoder and GFCL form the main structure of SES. CVA (Fig. 2) forms the main of VFS. The single arrow denotes the dataflow of SES and the double arrows denotes the dataflow of VFS.

In this paper, we introduce granular-ball to maximize the extraction of information structures from similar data, thereby extracting representations of mutual information. Compared to existing MVL paradigms, our method represents multi-view data with different quality using multi-granularity balls, which mitigates the influence of quality differences and encourages the model to extract more robust mutual representations. Our design maximizes the preservation of multi-view data with quality differences, rather than simply removing low-quality views (Su et al. 2025) or assigning weights for multiple views (Xu et al. 2023).

Methodology

Overview

As shown in Fig. 1, the proposed GFSAF method adopts a two-stage learning paradigm to train the model, consisting of SES and VFS, which is based on the Encoder-Decoder (E-D) framework to construct the network.

Given a multi-view database $\{X^i\}_{i=1}^V$ with M samples, each sample has multiple instances from V different views. In SES, each view is encoded by its corresponding encoder into a mutual representation $H^i \in \mathbb{R}^{b\times D_H}$ of consistent dimension across views. Then H^i is fed into multi-view GFCL, which progressively enhances the quality of H^i as training epochs iterate. Subsequently, we take H^i as the input of deeper fully connected layers to extract the latent complementary representations $Z^i \in \mathbb{R}^{b\times D_Z}$. In the extraction of Z^i, we design a novel Noise Stripping Loss to mitigate the influence of noise on Z^i. Eventually, the concatenated representation $[H^i, Z^i]$ is passed into the decoder, forming a complete training loop during SES. In VFS, we propose CVA to more effectively fuse H^i and Z^i, yielding a more comprehensive representation. To ensure consistency in the space of semantic representation and feature dimension, we further adopt H^i to guide the output of CVA, enhancing the convergence efficiency and clustering performance of our model.

Granular-ball Fuzzy Contrastive Learning

Multi-view Contrastive Learning (MCL) has emerged as a crucial research direction within the field of MVL. However, the traditional clustering paradigm (Xu et al. 2022; Liu et al. 2024; Zhang et al. 2025c), which only considers single and simplistic MCL or reconstruction during pre-training, has become a key factor contributing to RD. To this end, we enable the model to explore information representations from a multi-granularity perspective by introducing granular-balls, which blur fine-grained details while emphasizing robust decision boundary information effectively. Unlike traditional approaches that rely solely on pointwise cosine similarity as the optimization in MCL, we employ InfoNCE (Information Noise Contrastive Estimation) as the loss, which effectively extracts H^i cross multiple views from the mutual information theoretic perspective. Based on these ideas, we propose a novel Granular-ball Fuzzy Contrastive loss to guide mutual representation learning more effectively. That is:

$$\mathcal{L}^{i,j}_{InfoNCE} = -\mathbb{E}_{s^+\in\mathcal{P}_{GB}}[s^+ - log(e^{s^+} + \sum_{s^-}^{\mathcal{N}_{GB}} e^{s^-})], \quad (1)$$

where $\mathcal{P}_{GB}$ denote the set of positive granular-ball pairs and $\mathcal{N}_{GB}$ is the set of negative granular-ball pairs in the i,j-th views. $s^+(s^-)$ denotes the cosine distance between the representations of positive (negative) granular-ball pair. For multi-view, we adopt the granular-ball representation to express the point set of each view as granular-ball set.

Granular-ball from different views within the same sample is denoted as $\mathcal{P}_{GB}$. In the same view, when sample points overlap between two granular-balls, they are also classified as $\mathcal{P}_{GB}$ to consider potential false negatives. Dissimilar granular-ball is designated as $\mathcal{N}_{GB}$.

$$\mathcal{L}_{GFC} = \frac{1}{V(V-1)} \sum_{i=1}^{V} \sum_{j=1}^{V} \mathcal{L}_{InfoNCE}^{i,j} \quad (\forall i \neq j), \quad (2)$$

where V denotes the number of views. By minimizing $\mathcal{L}_{GFC}$ to maximize the mutual information cross views, H^i tends toward a more robust mutual information representation. All proofs are given in Appendix A.

Split-Extract Stage

We build the complete framework of GFSAF based on the self-supervised E-D, in which two types of structure, Autoencoder (AE) and Denoising Autoencoder (DAE), are selected as the backbone for SES. Specifically, for the i-th view, $E^i(\cdot;\theta^i)$ and $D^i(\cdot;\phi^i)$ denote its encoder and decoder, respectively. Firstly, $E^i(\cdot;\theta^i)$ performs embedding-based representation extraction on raw data for H^i. That is:

$$H^i = E^i(X^i;\theta^i), \quad (3)$$

where θ^i denotes the learnable parameters of $E^i(\cdot)$. To further enhance the robustness of H^i, we employ multi-granularity balls to represent each view, which fuzzifies the decision boundary of inter-class. Next, we configure positive and negative granular-ball pairs for multi-view GFCL.

Noise Stripping Loss. Based on H^i, we employ deeper fully connected layers to extract Z^i that are specific to each view. The core idea for extracting Z^i lies in the specific nature of each view itself. However, noise information also shares this same specificity, which has long been a significant challenge in MVL and is one of the main reasons that existing methods struggle to effectively disentangle and extract Z^i. From the perspective of definition, complementary information is inherently correlated with mutual information in a complex manner, whereas noise represents a type of information that actively interferes with the learning of mutual information. Recognizing the inherent duality between complementary and noise information, we design a novel Noise Stripping Loss $\mathcal{L}_{NS}$ that aims to strip noise representation $N^i \in \mathbb{R}^{b\times D_Z}$ in order to significantly extract Z^i. That is:

$$Z^i = ReLUs_Z^i(FCs_Z^i(H^i)), \quad (4)$$

$$N^i = ReLUs_N^i(FCs_N^i(H^i)), \quad (5)$$

where $*_Z^i$ and $*_N^i$ represent the layers of the network for complementary and noise information, respectively.

$$\mathcal{L}_{NS} = \frac{1}{2} \sum_{i=1}^{V} (cos(Z^i, N^i) + 1), \quad (6)$$

where $cos(\cdot)$ denotes the cosine similarity between two vectors. Eventually, we concatenate H^i and Z^i to obtain the fusion representation $F^i \in \mathbb{R}^{b\times D_F}$ of SES, which is fed into $D^i(\cdot;\phi^i)$ to complete the training loop. That is:

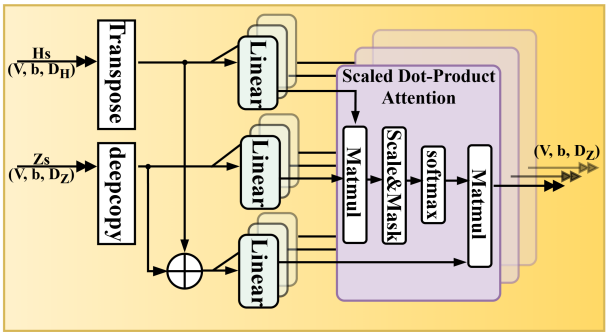


Figure 2: The proposed Cross Views Attention (CVA) module is as shown. More details can be found in VFS section.

$$F^i = concat(H^i, Z^i), \quad (7)$$

$$\mathcal{L}_{rec} = \frac{1}{V} \sum_{m=1}^{M} \sum_{i=1}^{V} ||X_m^i - D^i(F_m^i;\phi^i)||_2^2, \quad (8)$$

where $\mathcal{L}_{rec}$ denotes the reconstruction loss regularization term in SES. ϕ^i denotes the learnable parameters of $D^i(\cdot)$ and X_m^i denotes the m-th sample of X^i. In SES, we incorporate the reconstruction loss not as a primary optimization but rather as one regularization term, which encourages the model to place greater emphasis on the split and extraction of mutual and complementary information, rather than focusing on the reconstruction of the data.

Views-Fusion Stage

After the split and extraction of mutual and complementary representations in SES, our method proceeds to the next learning stage, which focuses on mining the complex nonlinear interdependence between H^i and Z^i, with the target of producing a more comprehensive fused representation $\hat{F}^i$.

Unlike existing multi-view fusion methods, we propose a novel CVA module, inspired by the attention mechanism, to effectively enhance the fusion between H^i and Z^i. As shown in Fig. 2, H and Z extracted for each view are concatenated separately to obtain the multi-view mutual representations Hs and complementary representations Zs, which then are fed into CVA. A transpose separation operation is applied to Hs to generate the key matrix F_K, while a deepcopy operation is applied to Zs to obtain the query matrix F_Q. The value matrix F_V is defined as a weighted mean of both Hs and Zs. Eventually, F_Q, F_K, and F_V are fed into CVA to generate the fused representation, which is formulated as:

$$F_V = mean(F_Q, F_K), \quad (9)$$

$$Attn(F_Q, F_K, F_V) = softmax(\frac{F_Q^T \times F_K}{\sqrt{D_Z}})F_V, \quad (10)$$

where $Attn(\cdot)$ denotes the attention mechanism. F_K represents the mutual information representation of the view to be matched for relevance, while F_Q denotes the complementary information representation used to query the relevant associations. The matrix F_V serves as the actual source of information, containing Hs and Zs, which are used for fusion.

Algorithm 1: Training steps of GFSAF

Input: Multi-view database $\{X^i\}_{i=1}^V$
Parameter: Batch size b. Hyper-parameters λ_1 and λ_2. Training epochs E_{ses}, E_{vfs}. Learning rate ℓ.
Output: The prediction $y^{'}$

for $e \in \{0, 1, ..., E_{ses} + E_{vfs}\}$ **do**
  for $l \in \{0, 1, ..., M/b\}$ **do**
    Pick mini-batch $\{X_m^i\}_{m=lb}^{(l+1)b}$ from $\{X^i\}_{i=1}^V$
    Compute the gradient of loss
    Update $\{\theta^i, \phi^i\}_{i=1}^V$ via Adam optimizer
    if $e < E_{ses}$ **then**
      $\theta^i = \theta^i - \frac{\ell}{b}\sum_{m=1}^{b} \frac{\partial \mathcal{L}_{SES}}{\partial \theta^i}$, and
      $\phi^i = \phi^i - \frac{\ell}{b}\sum_{m=1}^{b} \frac{\partial \mathcal{L}_{rec}}{\partial \phi^i}$
    end if
    if $e >= E_{ses}$ **then**
      $\theta^i = \theta^i - \frac{\ell}{b}\sum_{m=1}^{b} \frac{\partial \mathcal{L}_{rec}}{\partial \theta^i}$, and
      $\phi^i = \phi^i - \frac{\ell}{b}\sum_{m=1}^{b} \frac{\partial \mathcal{L}_{rec}}{\partial \phi^i}$
    end if
  end for
  Obtain $y^{'}$ by applying the K-means algorithm in $\hat{F}$
end for
return The prediction $y^{'}$

The interdependence between Hs and Zs is deeply mined via the operation of scaled dot-product, which generates an interdependence weight matrix used to guide the fusion. The introduction of multiple attention heads h enhances the ability of mining interdependence from multiple view perspectives. In CVA, we set the number of h equal to V, which encourages view-level attention focus and prevents attention distraction leading to overattention to details, caused by assigning more heads than views.

Eventually, the fused representation obtained via CVA and the H^i are concatenated for the final representation $\hat{F}^i$ in VFS. This design allows the model to retain connections to the original view representations, which serves for ensuring consistency in the space of semantic representation and feature dimension. The CVA module is not trained during the SES process and is initialized with random weights at the beginning of VFS. Directly using the output of CVA for decoding can result in a large difference of the semantic representation space between SES and VFS, potentially leading to high loss and unstable optimization. To this end, we employ mutual information as a form of semantic space guidance, which helps align the semantic distributions across stages, improves the convergence speed in VFS.

Joint Loss Function

In our GFSAF, SES and VFS are trained separately within the end-to-end manner. The joint loss functions for the two stages, denoted as $\mathcal{L}_{SES}$ and $\mathcal{L}_{VFS}$, are defined as follows:

$$\mathcal{L}_{SES} = \mathcal{L}_{GFC} + \lambda_1 \mathcal{L}_{NS} + \lambda_2 \mathcal{L}_{rec}, \tag{11}$$

$$\mathcal{L}_{VFS} = \mathcal{L}_{rec}, \tag{12}$$

where λ_1 and λ_2 denote the regularization parameters. By minimizing the loss in both stages, GFSAF effectively alleviates the RD, and extracts a fusion representation that better aligns with the fundamental objective of MVL.

Optimization. In the beginning, the parameters of GFSAF are randomly initialized. For unsupervised MVC task, we firstly utilize GFSAF and minimize $\mathcal{L}_{SES}$ in SES. Nextly, we employ $\mathcal{L}_{VFS}$ as the loss function in VFS. Specially, K-means is applied in $\hat{F}$ to obtain the clustering results. We utilize mini-batch gradient descent optimization to train the proposed GFSAF, which is summarized in Algorithm 1. More details of parameters can be found in next.

Experiments

Databases and Evaluation Setups

The databases involved in the experiments include **WebKB**, **Multi-COIL-10**, **Multi-COIL-20**, **Caltech101-7**, **Prokaryotic**, **NUSWIDE**, **Reuters**, **DHA**, and **UCI-Digits**, as shown in Table 1. The evaluation metric of Accuracy and PUR are selected. PUR, named pairwise unsupervised ranking, is one metric that measures consistency in MVC.

Database	Size	View	Class
WebKB (2007)	2,102	2	2-classes
Multi-COIL-10 (1996)	2,160	3	10-classes
Multi-COIL-20 (1996)	4,320	3	20-classes
Caltech101-7 (2004)	1,400	6	7-classes
Prokaryotic (2016)	1,653	3	4-classes
NUSWIDE (2010)	186,577	5	5-classes
Reuters (2009)	7,500	5	6-classes
DHA (2012)	966	2	23-classes
UCI-Digits (2007)	12,000	6	10-classes

Table 1: Detailed information of all benchmark databases.

Experimental Settings

The GFSAF method is implemented with the Pytorch toolbox, employing AE or DAE as the backbone, which is initialized with random weights.

The dimensions of H^i, Z^i and $\hat{F}^i$ are based on V. The value of D_Z is set to 64, the value of D_H is set to $V \times D_Z$ and D_F is set to $D_H + D_Z$. Based on extensive experiments in the databases, the values of λ_1 and λ_2 in Eq. (11) is empirically set to 0.3 and 0.1, respectively. We train GFSAF in an end-to-end manner with single NVIDIA GeForce RTX 4080 SUPER, and the batch size for all databases is set to 256. Then our model is trained using the Adam (Kingma and Ba 2014) with an initial learning rate of $5e-5$, weight decay = 0.01. More details are given in open-source code.

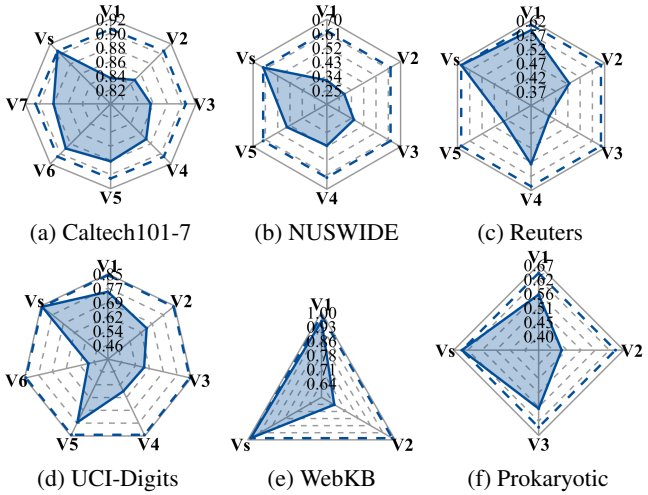
Ablation Studies

The ablation studies in this section are recorded in the databases. The ablation studies set up in this subsection focus on the challenge of alleviating the RD problem, the important modules of the proposed method (GFCL, CVA, $\mathcal{L}_{NS}$), and the weight values λ_1 and λ_2. The effect of each

Method	#Params	*WebKB*		*Multi-COIL-10*		*Multi-COIL-20*		*Caltech101-7*		*Prokaryotic*	
		Acc.↑	PUR↑	Acc.↑	PUR↑	Acc.↑	PUR↑	Acc.↑	PUR↑	Acc.↑	PUR↑
K-means (McQueen 1967)	−−	61.74	61.74	42.10	42.10	41.37	41.37	55.17	55.17	56.20	56.20
SEM (Xu et al. 2023)	20.17M	64.41	78.12	97.08	97.08	81.18	83.82	87.20	87.20	56.26	62.61
DealMVC (Yang et al. 2023)	25.12M	59.37	78.12	80.14	80.14	80.00	80.00	88.71	88.71	59.20	63.20
DIVIDE (Lu et al. 2024)	25.19M	88.01	89.12	98.16	98.16	80.39	81.24	62.20	63.51	54.99	54.99
FMCSC (Chen et al. 2024)	23.53M	54.80	54.80	93.75	93.75	82.80	82.80	61.57	61.57	54.16	54.16
RMCNC(Sun et al. 2024)	27.59M	81.59	81.59	89.32	90.57	80.96	81.28	69.16	69.16	54.83	56.19
MGBCC (Su et al. 2025)	21.35M	90.02	90.02	98.86	98.86	40.83	40.83	77.14	77.14	53.36	53.36
GFSAF$_{AE}$ (ours)	19.11M	**98.29**	**98.29**	**100.0**	**100.0**	83.26	84.17	88.74	88.74	**63.89**	76.59
GFSAF$_{DAE}$ (ours)	19.11M	94.13	94.13	**100.0**	**100.0**	**90.42**	**90.42**	**90.57**	**90.57**	62.07	**84.03**

Method	#Params	*NUSWIDE*		*Reuters*		*UCI-Digits*		*DHA*		*Average*	
		Acc.↑	PUR↑	Acc.↑	PUR↑	Acc.↑	PUR↑	Acc.↑	PUR↑	Acc.↑	PUR↑
K-means (McQueen 1967)	−−	39.73	40.26	31.27	31.27	45.26	45.26	65.60	65.60	48.71	48.77
SEM (Xu et al. 2023)	20.17M	60.60	60.60	56.50	56.50	79.64	79.64	**80.90**	**80.90**	72.85	75.37
DealMVC (Yang et al. 2023)	25.12M	25.92	25.96	47.05	48.36	73.16	73.16	48.65	48.65	61.39	64.07
DIVIDE (Lu et al. 2024)	25.19M	53.16	56.35	59.30	59.30	79.16	79.16	70.63	71.84	71.77	72.62
FMCSC (Chen et al. 2024)	23.53M	56.10	56.10	35.67	35.67	57.70	57.70	78.54	78.54	63.90	63.90
RMCNC(Sun et al. 2024)	27.59M	**67.59**	**67.59**	51.66	53.07	79.64	80.37	75.49	75.49	72.25	72.81
MGBCC (Su et al. 2025)	21.35M	27.94	30.47	41.83	44.83	46.40	47.60	68.22	68.22	60.51	61.26
GFSAF$_{AE}$ (ours)	19.11M	62.64	62.64	57.94	58.68	84.27	84.27	**80.90**	**80.90**	79.99	81.58
GFSAF$_{DAE}$ (ours)	19.11M	61.03	61.03	**60.50**	**60.50**	**88.50**	**88.50**	78.60	78.60	**80.65**	**83.09**

Table 2: Performance comparisons among different SoTA methods on several public multi-view databases. The best results are boldfaced and the second results are underlined. Acc.↑ denotes the accuracy (%) of performance in the database.



(a) Caltech101-7 (b) NUSWIDE (c) Reuters

(d) UCI-Digits (e) WebKB (f) Prokaryotic

Figure 3: Ablation of alleviating the problem of RD. v_i denotes the performance of single view and v_s denotes the fusion views. More results are shown in Appendix B.

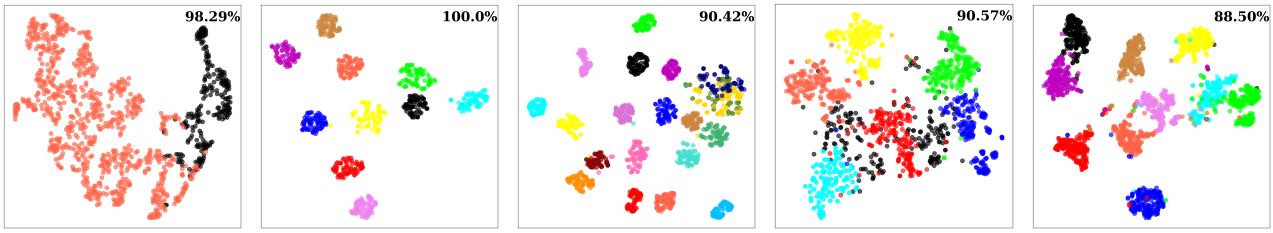
#1	#2	#3	*Acc.*↑ / *PUR*↑		
			WebKB	Caltech101-7	Prokaryotic
✗	✗	✗	88.64/88.64	77.89/77.89	59.47/59.47
✓	✗	✗	91.24/91.24	79.62/79.62	60.34/60.34
✗	✓	✗	89.96/89.96	78.12/78.12	59.98/59.98
✗	✗	✓	91.89/91.89	81.34/81.34	61.24/61.24
✓	✓	✗	94.03/94.03	84.86/84.86	62.67/74.07
✓	✗	✓	96.41/96.41	85.74/86.09	62.38/71.29
✗	✓	✓	93.46/93.46	84.37/84.37	62.97/74.79
✓	✓	✓	**98.29/98.29**	**88.74/88.74**	**63.89/76.59**

Table 3: The ablation study of the proposed key modules. #1, #2 and #3 denote the GFCL, $\mathcal{L}_{NS}$ and CVA respectively.

module in the proposed method is explored through ablation studies on the databases. Ablation studies for the other model parameters are shown in Appendix B.

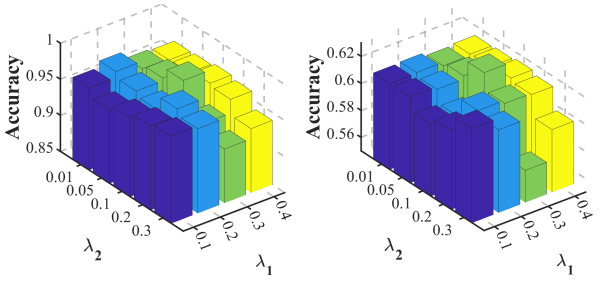
Challenge of alleviating the RD problem. We show our GFSAF method about RD existing in current MVL paradigm. As shown in Fig. 3, we can observe that the performance of $\hat{F}$ in our method is significantly higher than that of any other single-view representation. In particular, in the Caltech101-7, NUSWIDE, and Prokaryotic databases, the accuracy performance of $\hat{F}$ is on average 3.97%, 25.98% and 11.92% superior than that of v_i, respectively. In all databases, our method shows no any degradation in term of representation quality, which sufficiently demonstrates that the novel MVL paradigm we propose effectively deals with this existing challenge. The proposed paradigm does not merely aim to improve fusion representation performance, but also balances the extraction of effective information from other views and the effective fusion of multi-view information, which aligns more closely with the core of MVL.

Influence of the key modules. As shown in Table 3, when only a single key module is employed in our model, the accuracy performance in three databases all shows a slight improvement. The primary reason is that the three key modules are all designed form the core of MVL, improving information extraction or enhancing fusion, to enhance the ability of



(a) WebKB (b) Multi-COIL-10 (c) Multi-COIL-20 (d) Caltech101-7 (e) UCI-Digits

Figure 4: Visualization of fusion features $\hat{F}$ using t-SNE on five databases.



(a) Influence in WebKB (b) Influence in NUSWIDE

Figure 5: Ablation studies for the different of λ_1 and λ_2.

GFSAF. It should be noted that the improvement achieved by using CVA alone is smaller compared to that of other modules. The primary reason is that the CVA is designed to enhance the fusion effect of the model, which depends on the information quality of H^i and Z^i. In particular, when the other two modules are used in combination with CVA, the accuracy performance improves significantly, with the average final performance being 3.12% superior.

Influence of the λ_1 and λ_2. We evaluate the accuracy performance of our method with the different weight values of λ_1 and λ_2 in Eq. (11), as shown in Fig. 5. In particular, we can observe that GFSAF achieves the best accuracy when the values of λ_1 and λ_2 are set to 0.3 and 0.1, respectively.

When a large value of λ_1 is set, GFSAF tends to focus excessively on the separation of Z^i and N^i, which is theoretically achievable. However, $\mathcal{L}_{NS}$ cannot strictly be regarded as the separation of noise. Furthermore, $\mathcal{L}_{NS}$ represents an approximate fitting of the separation loss. The relationship between N^i and Z^i is complex, and $\mathcal{L}_{NS}$ certainly belongs to this relationship, but it cannot fully represent it. Therefore, excessive focus on $\mathcal{L}_{NS}$ is not beneficial for model learning. In addition, when a large value of λ_2 is set, GFSAF may overly focus the view reconstruction, which is not beneficial to learning multi-view information. A lower value of $\mathcal{L}_{rec}$ only indicates that the hidden layer representations effectively represent the view data, but this representation has drawbacks in MVL. Under this learning paradigm, there is excessive similarity between the hidden representations of multiple views, which is not conducive to the fusion, which is particularly easy to causing the RD problem.

Comparisons with State-of-The-Art Methods

Table 2 shows the comparison results between the proposed GFSAF method and several state-of-the-art methods on all databases shown in Table 1. Only a different portion of the databases in Table 1 is generally selected in the article for each comparison method. To ensure more complete and effective comparability, we use the open source code of the comparison methods to conduct several experiments on the databases without records to obtain the optimal results, which are then compared with our method.

As shown in Table 2, our GFSAF method basically improves the accuracy and PUR performance on all databases with the average improvement of 6.58% and 7.72%, respectively. The accuracy is significantly improved by 8.27% and 8.86% in WebKB and UCI-Digits, respectively. It also improves on all other databases except NUSWIDE. These comparison experiments with the state-of-the-art method clearly demonstrate our advantage in MVC as well as the existence of the proposed GFSAF method with high robustness on different databases. In addition, the method we propose has lighter parameters, which indirectly proves that the improved performance of our method does not necessarily depend on the complex backbone. Our method focuses more on creating a novel paradigm to deal with the common problems of current MVL methods and improve the performance of MVL models by dealing with core challenges. This learning paradigm can also be transferred to other MVL tasks.

2D feature visualization. We use t-SNE (Maaten and Hinton 2008) to visualize $\hat{F}$ of GFSAF on different databases on the 2D space, respectively, as shown in Fig. 4. We can conclude that $\hat{F}$ extracted by GFSAF has good classification results not only works well within the network.

Conclusion

In this paper, we propose a novel GFSAF consisting of SES and VFS. To deal with the challenges of quality discrepancy and RD broadly existing in current methods, we import granular-ball into MVL and design the GFCL and $\mathcal{L}_{NS}$, which obtain mutual representation and complementary representation effectively via the Split-Extract design. In addition, we innovatively introduce attention mechanism and propose CVA, an attention variant, for more effective representations fusion to alleviate the RD. We conduct ablation studies to demonstrate the contributions we mentioned.

Acknowledgments

This work is supported in part by National Natural Science Foundation of China (Nos. 62221005, 62476052, 62450043, and 62222601), Radiation Oncology Key Laboratory of Sichuan Province Open Fund (No. 2024ROKF05), and the Open Fund of the Key Laboratory of Cyberspace Big Data Intelligent Security, Ministry of Education (No. CBDIS202501).

References

Amini, M. R.; Usunier, N.; and Goutte, C. 2009. Learning from multiple partially observed views-an application to multilingual text categorization. *NeurIPS*, 22.

Asuncion, A.; Newman, D.; et al. 2007. UCI machine learning repository.

Brbić, M.; Piškorec, M.; Vidulin, V.; Kriško, A.; Šmuc, T.; and Supek, F. 2016. The landscape of microbial phenotypic traits and associated genes. *NUCLEIC ACIDS RES*, gkw964.

Chen, L. 1982. Topological structure in visual perception. *Science*, 218(4573): 699–700.

Chen, X.; Ren, Y.; Xu, J.; Lin, F.; Pu, X.; and Yang, Y. 2024. Bridging gaps: Federated multi-view clustering in heterogeneous hybrid views. *NeurIPS*, 37: 37020–37049.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR*, 178–178.

He, H.; Xu, J.; Wen, G.; Ren, Y.; Zhao, N.; and Zhu, X. 2025. Graph Embedded Contrastive Learning for Multi-View Clustering. In *IJCAI*, 5336–5344.

Hu, L.; Chen, F.; Zhao, S.; Duan, S.; et al. 2025. GRICP: Granular-Ball Iterative Closest Point with Multikernel Correntropy for Point Cloud Fine Registration. In *AAAI*, volume 39, 1710–1718.

Huang, J.; Cheung, Y.-m.; Vong, C.-m.; and Qian, W. 2025. GBRIP: Granular Ball Representation for Imbalanced Partial Label Learning. In *AAAI*, volume 39, 17431–17439.

Jiang, X.; He, B.; Zhou, P. Y.; Chen, X.; Guo, J.; Xu, J.; and Liao, Y. 2025. A Unified Framework to BRIDGE Complete and Incomplete Deep Multi-View Clustering under Non-IID Missing Patterns. In *ICCV*, 594–603.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, Y.; Ouyang, X.; Pan, C.; Zhang, J.; Zhao, S.; Xia, S.; Yang, X.; Wang, G.; and Li, T. 2025. Multi-Granularity Open Intent Classification via Adaptive Granular-Ball Decision Boundary. In *AAAI*, volume 39, 24512–24520.

Li, Y.; Zhang, D.; Yang, M.; Peng, D.; Yu, J.; Liu, Y.; Lv, J.; Chen, L.; and Peng, X. 2023. scBridge embraces cell heterogeneity in single-cell RNA-seq and ATAC-seq data integration. *Nat. Commun.*, 14(1): 6045.

Lin, Y.-C.; Hu, M.-C.; Cheng, W.-H.; Hsieh, Y.-H.; and Chen, H.-M. 2012. Human action recognition and retrieval using sole depth information. In *ACM MM*, 1053–1056.

Liu, S.; Liang, K.; Dong, Z.; Wang, S.; Yang, X.; Zhou, S.; Zhu, E.; and Liu, X. 2024. Learn from view correlation: An anchor enhancement strategy for multi-view clustering. In *CVPR*, 26151–26161.

Lu, F.; Hou, Y.; Li, W.; Yang, X.; Zheng, H.; Luo, W.; Chen, L.; Cao, Y.; Liao, X.; Zhang, Y.; et al. 2025. NaFV-Net: An Adversarial Four-view Network for Mammogram Classification. In *AAAI*, volume 39, 28213–28221.

Lu, Y.; Lin, Y.; Yang, M.; Peng, D.; Hu, P.; and Peng, X. 2024. Decoupled contrastive multi-view clustering with high-order random walks. In *AAAI*, volume 38, 14193–14201.

Luo, C.; Xu, J.; Ren, Y.; Ma, J.; and Zhu, X. 2024. Simple Contrastive Multi-View Clustering with Data-Level Fusion. In *IJCAI*, 4697–4705.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *J Mach Learn Res.*, 9(Nov): 2579–2605.

McAllester, D.; and Stratos, K. 2020. Formal limitations on the measurement of mutual information. In *AISTATS*, 875–884.

McQueen, J. B. 1967. Some methods of classification and analysis of multivariate observations. In *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.*, 281–297.

Nene, S. A.; Nayar, S. K.; Murase, H.; et al. 1996. Columbia object image library (coil-20).

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Quadir, A.; and Tanveer, M. 2024. Granular ball twin support vector machine with pinball loss function. *IEEE TCSS*.

Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G. R.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *ACM MM*, 251–260.

Ren, Y.; Chen, X.; Xu, J.; Pu, J.; Huang, Y.; Pu, X.; Zhu, C.; Zhu, X.; Hao, Z.; and He, L. 2024a. A novel federated multi-view clustering method for unaligned and incomplete data fusion. *INFORM FUSION*, 108: 102357.

Ren, Y.; Ke, J.; Wen, Z.; Wu, T.; Yang, Y.; Pu, X.; and He, L. 2025. Multi-View Graph Clustering via Node-Guided Contrastive Encoding. In *ICML*.

Ren, Y.; Pu, J.; Cui, C.; Zheng, Y.; Chen, X.; Pu, X.; and He, L. 2024b. Dynamic weighted graph fusion for deep multi-view clustering. In *IJCAI*, 4842–4850.

Su, P.; Huang, S.; Ma, W.; Xiong, D.; and Lv, J. 2025. Multi-view Granular-ball Contrastive Clustering. In *AAAI*, volume 39, 20637–20645.

Sun, T.-K.; Chen, S.-C.; Jin, Z.; and Yang, J.-Y. 2007. Kernelized discriminative canonical correlation analysis. In *ICWAPR*, volume 3, 1283–1287.

Sun, Y.; Qin, Y.; Li, Y.; Peng, D.; Peng, X.; and Hu, P. 2024. Robust multi-view clustering with noisy correspondence. *IEEE TKDE*.

Sun, Y.; Ren, Z.; Hu, P.; Peng, D.; and Wang, X. 2023. Hierarchical consensus hashing for cross-modal retrieval. *IEEE TMM*, 26: 824–836.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In *ECCV*, 776–794.

Wang, T.; and Isola, P. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 9929–9939. PMLR.

Wu, H.; Gattami, A.; and Flierl, M. 2020. Conditional mutual information-based contrastive loss for financial time series forecasting. In *ICAIF*, 1–7.

Wu, S.; Zheng, Y.; Ren, Y.; He, J.; Pu, X.; Huang, S.; Hao, Z.; and He, L. 2024. Self-weighted contrastive fusion for deep multi-view clustering. *IEEE TMM*, 26: 9150–9162.

Xia, S.; Dai, D.; Chen, F.; Yang, L.; Wang, G.; Wang, G.; and Gao, X. 2025a. An Adaptive Multi-Granularity Graph Representation of Image via Granular-Ball Computing. *IEEE TIP*.

Xia, S.; Liu, Y.; Ding, X.; Wang, G.; Yu, H.; and Luo, Y. 2019. Granular ball computing classifiers for efficient, scalable and robust learning. *Information Sciences*, 483: 136–152.

Xia, S.; Ma, X.; Liu, Z.; Liu, C.; Zhao, S.; and Wang, G. 2025b. Graph coarsening via supervised granular-ball for scalable graph neural network training. In *AAAI*, volume 39, 12872–12880.

Xie, J.; Jiang, L.; Xia, S.; Xiang, X.; and Wang, G. 2024. An adaptive density clustering approach with multi-granularity fusion. *INFORM FUSION*, 106: 102273.

Xu, J.; Chen, S.; Ren, Y.; Shi, X.; Shen, H.; Niu, G.; and Zhu, X. 2023. Self-weighted contrastive learning among multiple views for mitigating representation degeneration. *NeurIPS*, 36: 1119–1131.

Xu, J.; Tang, H.; Ren, Y.; Peng, L.; Zhu, X.; and He, L. 2022. Multi-level feature learning for contrastive multi-view clustering. In *CVPR*, 16051–16060.

Xu, J.; Zhao, N.; Niu, G.; Sugiyama, M.; and Zhu, X. 2025. Robust Multi-View Learning via Representation Fusion of Sample-Level Attention and Alignment of Simulated Perturbation. In *ICCV*, 4232–4241.

Yang, X.; Jiaqi, J.; Wang, S.; Liang, K.; Liu, Y.; Wen, Y.; Liu, S.; Zhou, S.; Liu, X.; and Zhu, E. 2023. Dealmvc: Dual contrastive calibration for multi-view clustering. In *ACM MM*, 337–346.

Zhang, C.; Cui, Y.; Han, Z.; Zhou, J. T.; Fu, H.; and Hu, Q. 2020. Deep partial multi-view learning. *IEEE TPAMI*, 44(5): 2402–2415.

Zhang, H.; Yue, H.; Xiao, X.; Yu, L.; Li, Q.; Ling, Z.; and Zhang, Y. 2025a. Revolutionizing Encrypted Traffic Classification with MH-Net: A Multi-View Heterogeneous Graph Model. In *AAAI*, volume 39, 1048–1056.

Zhang, Y.; Lin, Y.; Yan, W.; Yao, L.; Wan, X.; Li, G.; Zhang, C.; Ke, G.; and Xu, J. 2025b. Incomplete Multi-view Clustering via Diffusion Contrastive Generation. In *AAAI*, volume 39, 22650–22658.

Zhang, Y.; Yan, W.; Tang, C.; Zhou, W.; and Jin, J. 2025c. Multi-branch Space Sharing Feature Aggregation for contrastive multi-view clustering. *Pattern Recognition*, 111704.

Zhong, H.; Chen, C.; Jin, Z.; and Hua, X.-S. 2020. Deep robust clustering by contrastive learning. *arXiv preprint arXiv:2008.03030*.

Appendix. We provide supplementary materials for the submission of *Views Attention Fusion of Granular-ball Fuzzy Representations Split for Improving Multi-view Clustering*. Here are the contents:

- Appendix A: Theoretical Analysis.
- Appendix B: Additional Experiment Results.

A. Theoretical Analysis

In this section, we show more details about all theoretical analysis of GFSAF.

Theorem 1. *For any two views $(i, j \in \{1, 2, ..., V\})$ with positive class mutual information $I(y^i, y^j) = \delta, \delta > 0$, if E^i learned by GFCL is a smooth invertible transformation, minimizing the Granular-ball Fuzzy Contrastive loss $\mathcal{L}_{GFC}$ will lead to a trade-off between max $I(X^i; X^j; H^i; H^j)$.*

Proof. According to Proposition 1, minimizing the joint loss approximately becomes maximizing the following objective:

$$\begin{aligned}\mathcal{L}_{GFC} &= \mathcal{L}^{i,j}_{InfoNCE}(H^i_{GB}, H^j_{GB}) \\ &= (e^{\delta/logM} - 1)I(H^i_{GB}; H^j_{GB}).\end{aligned} \tag{13}$$

If transformations E^i and E^j are smooth and invertible, the Jacobian determinant is $J_{H^i_{GB}} = |\frac{\partial H^i_{GB}}{\partial H^i}|$ and $J_{H^j_{GB}} = |\frac{\partial H^j_{GB}}{\partial H^j}|$, respectively. For the i-th and j-th views, we have:

$$\begin{aligned}p(h^i, h^j) = p(h^i_{GB}, h^j_{GB}) J_{H^i_{GB}}(h^i) J_{H^j_{GB}}(h^j), \\ p(h^i) = p(h^i_{GB}) J_{H^i_{GB}}(h^i), dh^i_{GB} = J_{H^i_{GB}}(h^i_{GB}) dh^i, \\ p(h^j) = p(h^j_{GB}) J_{H^j_{GB}}(h^j), dh^j_{GB} = J_{H^j_{GB}}(h^j_{GB}) dh^j.\end{aligned} \tag{14}$$

Then, we can obtain the invariance property of mutual information between $I(H^i_{GB}; H^j_{GB})$ and $I(H^i; H^j)$ as follows:

As a result, the optimization objective in Eq. (13) becomes:

$$\mathcal{L}_{GFC} = (e^{\delta/logM} - 1)I(H^i; H^j). \tag{15}$$

The mutual information $I(X^i; X^j)$ in X^i and X^j is fixed, and the mutual information $I(H^i; H^j)$ changes due to variables H^i and H^j. Maximizing $I(H^i; H^j)$ tends to make variables to access $I(X^i; X^j)$.

Proposition 1. *Minimizing the GFCL InfoNCE loss $\sum \mathcal{L}^{i,j}_{InfoNCE}(H^i_{GB}, H^j_{GB})$ among the representations of multiple views is equivalent to maximizing the mutual mutual information $\sum I(H^i_{GB}, H^j_{GB})$.*

Proof. In the part, we leverage $d(h^{'i}_m, h^{'j}_m)$ to denote the cosine distance between $h^{'i}_m \in H^i_{GB}$ and $h^{'j}_m \in H^j_{GB}$. Then, based on the inequality in Lemma 1, we have:

$$-\frac{1}{M}\sum_{m=1}^{M} log(\frac{e^{d(h^{'i}_m, h^{'j}_m)/\tau}}{\sum_{l=1}^{M} e^{d(h^{'i}_m, h^{'j}_l)}/\tau}) \geq logM - I(H^i_{GB}; H^j_{GB}) \tag{16}$$

We rewrite the positive and negative pairs in InfoNCE loss and can obtain the following inequality:

$$\begin{aligned}&-\frac{1}{M}\sum_{m=1}^{M} log(\frac{e^{d(h^{'i}_m, h^{'j}_m)/\tau}}{\sum_{l=1}^{M}\sum_{v=i,j} e^{d(h^{'i}_m, h^{'v}_l)}/\tau}) \\ &\geq -\frac{1}{M}\sum_{m=1}^{M} log(\frac{e^{d(h^{'i}_m, h^{'j}_m)/\tau}}{\sum_{l=1}^{M} e^{d(h^{'i}_m, h^{'j}_l)}/\tau}) \\ &\geq logM - I(H^i_{GB}; H^j_{GB})\end{aligned} \tag{17}$$

Given the equations $I(H^i_{GB}; H^j_{GB}) = I(H^j_{GB}; H^i_{GB})$, we further have:

$$\begin{aligned}&\sum_{i,j}^{V} \mathcal{L}^{i,j}_{InfoNCE}(H^j_{GB}; H^i_{GB}) \\ &\geq \sum_{i=1}^{V}\sum_{j=1}^{V}(logM - I(H^i_{GB}; H^j_{GB})) \\ &= V^2 logM - 2\sum_{i=1}^{V}\sum_{j=1}^{V} I(H^i_{GB}; H^j_{GB})\end{aligned} \tag{18}$$

Therefore, min $\sum_{i,j}^{V} \mathcal{L}^{i,j}_{InfoNCE}(H^j_{GB}; H^i_{GB})$ is equivalent to max $\sum_{i,j}^{V} I(H^i_{GB}; H^j_{GB}), i.e.$ minimizing the GFCL InfoNCE loss among the representations of multi-view is equivalent to maximizing the mutual information.

The success of contrastive learning is often (not absolutely) attributable to the estimation of mutual information. Eq. (19) gives the relation between InfoNCE and mutual information, which also has been discussed by other forms in (Tian, Krishnan, and Isola 2020; Wu, Gattami, and Flierl 2020; Wang and Isola 2020; Oord, Li, and Vinyals 2018; Zhong et al. 2020). In our paper, we rewrite proofs for this inequality for the completeness of lemmas.

Lemma 1. *Let i and j denotes two views, assuming $p()$ assuming $p(h^{'i}_m, h^{'j}_l) = p(h^{'i}_m)p(h^{'j}_l)$ when $m \neq l$, we have the following inequality that give the relation between InfoNCE and mutual information:*

$$\begin{aligned}&-\frac{1}{M}\sum_{m=1}^{M} log(\frac{exp(d(h^{'i}_m, h^{'j}_m)/\tau)}{\sum_{l=1}^{M} exp(d(h^{'i}_m, h^{'j}_l)/\tau)}) \\ &\geq logM - I(H^i_{GB}; H^j_{GB}).\end{aligned} \tag{19}$$

Proof. If $m \neq l, p(h^{'j}_l | h^{'i}_m) = \frac{p(h^{'j}_l, h^{'i}_m)}{p(h^{'i}_m)} = p(h^{'j}_l)$. Let

$\mathcal{S}_m = \sum_{l=1}^{M} \frac{p(h_l^{'j}, h_m^{'i})}{p(h_l^{'j})p(h_m^{'i})}$, then, we gain:

$$
\begin{aligned}
I(H_{GB}^i; H_{GB}^j) &= \sum_{m=1}^{M}\sum_{l=1}^{M} p(h_m^{'i}, h_l^{'j}) log \frac{p(h_m^{'i}, h_l^{'j})}{p(h_m^{'i})p(h_l^{'j})} \\
&= \sum_{m=1}^{M}\sum_{l=1}^{M} p(h_m^{'i}, h_l^{'j}) log (\frac{p(h_m^{'i}, h_l^{'j})\mathcal{S}_m}{p(h_m^{'i})p(h_l^{'j})\mathcal{S}_m}) \\
&= \sum_{m=1}^{M}\sum_{l=1}^{M} p(h_m^{'i}, h_l^{'j}) log \frac{\frac{p(h_m^{'i}, h_l^{'j})}{p(h_m^{'i})p(h_l^{'j})}}{\mathcal{S}_m} \\
&+ \sum_{m=1}^{M}\sum_{l=1}^{M} p(h_m^{'i}, h_l^{'j}) log \mathcal{S}_m \\
&= \sum_{m=1}^{M} p(h_m^{'i}, h_m^{'j}) log \frac{\frac{p(h_m^{'i}, h_m^{'j})}{p(h_m^{'i})p(h_m^{'j})}}{\mathcal{S}_m} \\
&+ \sum_{m=1}^{M}\sum_{m\neq l} p(h_m^{'i}, h_l^{'j}) log \frac{\frac{p(h_m^{'i}, h_l^{'j})}{p(h_m^{'i})p(h_l^{'j})}}{\mathcal{S}_m} \\
&+ \sum_{m=1}^{M}\sum_{l=1}^{M} p(h_m^{'i}, h_l^{'j}) log \mathcal{S}_m \\
&= \sum_{m=1}^{M} p(h_m^{'i}, h_m^{'j}) log \frac{\frac{p(h_m^{'i}, h_m^{'j})}{p(h_m^{'i})p(h_m^{'j})}}{\mathcal{S}_m} \\
&+ \sum_{m=1}^{M}\sum_{m\neq l} p(h_m^{'i}, h_l^{'j}) log \frac{p(h_m^{'i}, h_l^{'j})}{p(h_m^{'i})p(h_l^{'j})} \\
&+ \sum_{m=1}^{M}\sum_{l=1}^{M} p(h_m^{'i}, h_l^{'j}) log \mathcal{S}_m \\
&- \sum_{m=1}^{M}\sum_{m\neq l} p(h_m^{'i}, h_l^{'j}) log \mathcal{S}_i \\
&= \sum_{m=1}^{M} p(h_m^{'i}, h_m^{'j}) log \frac{\frac{p(h_m^{'i}, h_m^{'j})}{p(h_m^{'i})p(h_m^{'j})}}{\mathcal{S}_m} \\
&+ \sum_{m=1}^{M} p(h_m^{'i}, h_m^{'j}) log \mathcal{S}_m.
\end{aligned}
\tag{20}
$$

Since positive pairs are correlated, we have the estimate: $p(h_m^{'i}, h_m^{'j}) \geq h_m^{'j} p(h_m^{'i})$. Therefore, the following inequality holds:

$$
\begin{aligned}
log\mathcal{S}_m &= log(\sum_{l=1}^{M} \frac{p(h_m^{'i}, h_l^{'j})}{p(h_m^{'i})p(h_l^{'j})}) \\
&= log(\frac{p(h_m^{'i}, h_m^{'j})}{p(h_m^{'i})p(h_m^{'j})} + \sum_{m\neq l} \frac{p(h_m^{'i}, h_l^{'j})}{p(h_m^{'i})p(h_l^{'j})}) \\
&= log(M + \frac{p(h_m^{'i}, h_m^{'j})}{p(h_m^{'i})p(h_m^{'j})} - 1) \\
&\geq logN.
\end{aligned}
\tag{21}
$$

According to Lemma 2 and Eq. (21), we assume that there exists a constant $\delta \in (0, 1)$ such that $p(h_m^{'i}|h_m^{'j}) \neq \delta, m = 1, 2, ..., M$ holds. With the estimation (Oord, Li, and Vinyals 2018; Zhong et al. 2020), *i.e.*, $p(h_m^{'j}) \approx \frac{1}{M}, m = 1, 2, ..., M$, the following inequality holds:

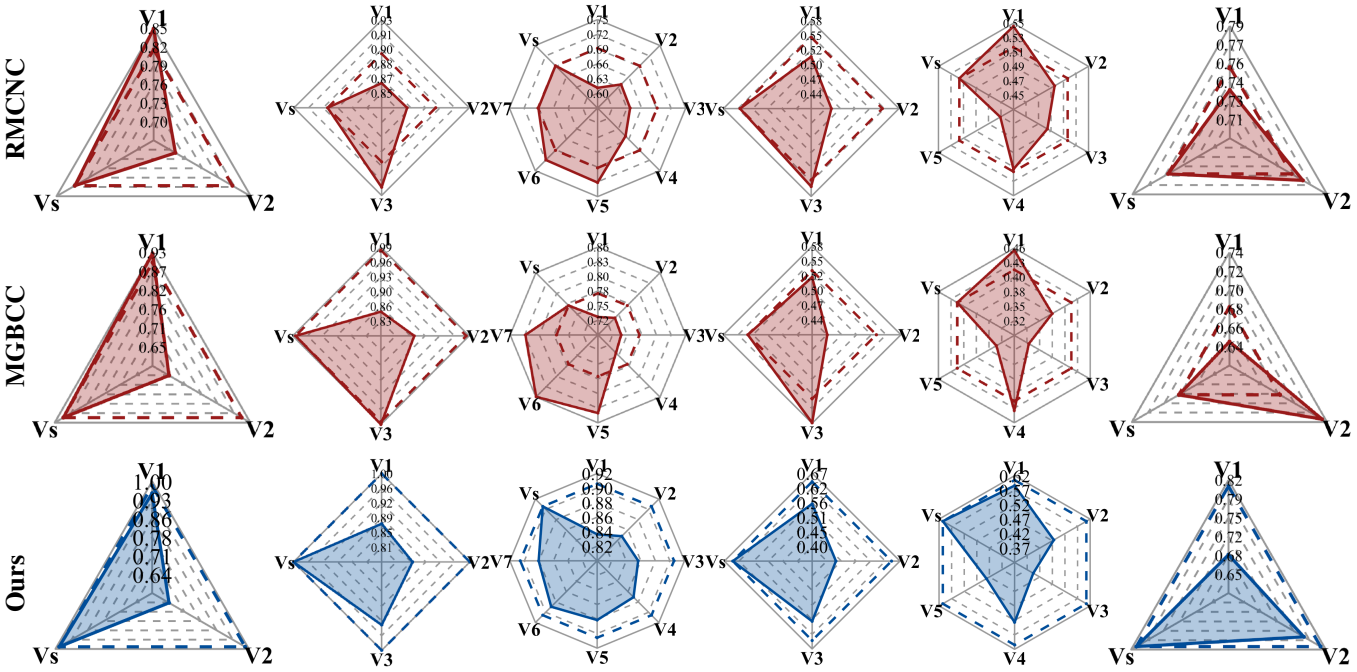
$$
\begin{aligned}
I(H_{GB}^i; H_{GB}^j) &= \sum_{m=1}^{M} p(h_m^{'i}, h_m^{'j}) log \frac{\frac{p(h_m^{'i}, h_m^{'j})}{p(h_m^{'i})p(h_m^{'j})}}{\mathcal{S}_m} \\
&+ \sum_{m=1}^{M} p(h_m^{'i}, h_m^{'j}) log \mathcal{S}_m \\
&\approx \sum_{m=1}^{M} \frac{1}{M} p(h_m^{'i}|h_m^{'j}) log \frac{\frac{p(h_m^{'i}, h_m^{'j})}{p(h_m^{'i})p(h_m^{'j})}}{\mathcal{S}_m} \\
&+ \sum_{m=1}^{M} \frac{1}{M} p(h_m^{'i}|h_m^{'j}) log \mathcal{S}_m \\
&\geq \delta(\frac{1}{M}\sum_{m=1}^{M} log \frac{exp(d(h_m^{'i}, h_m^{'j}))/\tau}{\sum_{l=1}^{M} exp(d(h_m^{'i}, h_l^{'j}))/\tau} \\
&+ logN).
\end{aligned}
\tag{22}
$$

Furthermore, we can gain:

$$
\begin{aligned}
-\frac{1}{M}\sum_{m=1}^{M} log \frac{exp(d(h_m^{'i}, h_m^{'j}))/\tau}{\sum_{l=1}^{M} exp(d(h_m^{'i}, h_l^{'j}))/\tau} &\geq logN \\
&- \frac{1}{\delta} I(H_{GB}^i; H_{GB}^j).
\end{aligned}
\tag{23}
$$

Consequently, when the constant $\delta \approx 1$ (*i.e.*, the positive pairs are approximate to be correlated), Eq. (19) holds. According to (Oord, Li, and Vinyals 2018), Eq. (23) is more precise when M is larger. Minimizing the left part of Eq. (23) is equivalent to maximizing the mutual information $I(H_{GB}^i; H_{GB}^j)$. Note that this bound is weak as there exists approximation about mutual information (McAllester and Stratos 2020).

Lemma 2. *The optimal value of* $exp(d(h_m^{'i}, h_l^{'j}))/\tau$ *is proportional to the ratio of* $p(h_m^{'i}, h_l^{'j})$ *to* $p(h_m^{'i})p(h_l^{'j})$*, i.e.,* $exp(d(h_m^{'i}, h_l^{'j}))/\tau \propto \frac{p(h_m^{'i}, h_l^{'j})}{p(h_m^{'i})p(h_l^{'j})}$.



(a) WebKB (b) Multi-COIL-10 (c) Caltech101-7 (d) Prokaryotic (e) Retuers (f) DHA

Figure 6: More ablation experiments of alleviating the problem of RD. v_i denotes the performance of single view and v_s denotes the fusion views. The upper shows the performance of one selected SoTA comparison method (RMCNC) in RD. The middle shows the performance of one selected SoTA comparison method (MGBCC) in RD. The lower shows the performance of our method in RD.

Proof. We consider the following formulation:

$$-\frac{1}{M}\sum_{m=1}^{M} log\frac{exp(d(h_m^{'i},h_m^{'j}))/\tau}{\sum_{l=1}^{M} exp(d(h_m^{'i},h_l^{'j}))/\tau}. \tag{24}$$

Eq. (24) can be regarded as a cross-entropy loss. As a result, minimizing this loss is equivalent to solve a binary classification problem, namely, classifying the given pairs into positive or negative pairs. We let $\{h_m^{'i},h_m^{'j}\}$ denote the positive pairs and $\{h_m^{'i},h_l^{'j}\},(m \neq l)$ denote the negative pairs. For each given pairs $\{h_m^{'i},h_m^{'j}\}$, we let $p(h_m^{'j}|\{h_1^{'j},h_2^{'j},...,h_M^{'j}\},h_m^{'i})$ denote the predicted probability of finding $h_m^{'j}$ from $\{h_1^{'j},h_2^{'j},...,h_M^{'j}\}$ to form positive pairs $\{h_m^{'i},h_m^{'j}\}$. $p(h_m^{'i},h_l^{'j}),p(h_m^{'i})$ and $p(h_l^{'j})$ denote the joint probability and marginal probabilities of $h_m^{'i}$ and $h_l^{'j}$. Then the optimal value of $p(h_m^{'j}|\{h_1^{'j},h_2^{'j},...,h_M^{'j}\},h_m^{'i})$ is:

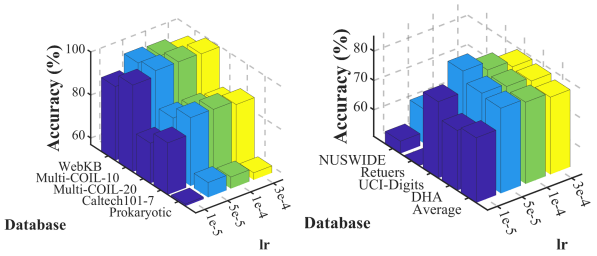
$$\begin{aligned} p(h_m^{'j}|\{h_1^{'j},h_2^{'j},...,h_M^{'j}\},h_m^{'i}) &= \frac{p(h_m^{'j}|h_m^{'i})\pi_{k\neq m}p(h_k^{'j})}{\sum_{l=1}^{M}p(h_l^{'j}|h_m^{'i})\pi_{k\neq l}p(h_k^{'j}} \\ &= \frac{p(h_m^{'j}|h_m^{'i})}{p(h_m^{'j})}\sum_{l=1}^{M}\frac{p(h_l^{'j}|h_m^{'i})}{p(h_l^{'j})} \\ &= \frac{\frac{p(h_m^{'j},h_m^{'i})}{p(h_m^{'i})p(h_m^{'j})}}{\sum_{l=1}^{M}\frac{p(h_l^{'j},h_m^{'i})}{p(h_m^{'i})p(h_l^{'j})}}. \end{aligned} \tag{25}$$

The corresponding cross-entropy loss is:

$$\begin{aligned} \mathcal{L} &= -\frac{1}{M}\sum_{m=1}^{M} logp(h_m^{'j}|\{h_1^{'j},h_2^{'j},...,h_M^{'j}\},h_m^{'i}) \\ &= -\frac{1}{M}\sum_{m=1}^{M} log\frac{\frac{p(h_m^{'j},h_m^{'i})}{p(h_m^{'i})p(h_m^{'j})}}{\sum_{l=1}^{M}\frac{p(h_l^{'j},h_m^{'i})}{p(h_m^{'i})p(h_l^{'j})}}. \end{aligned} \tag{26}$$

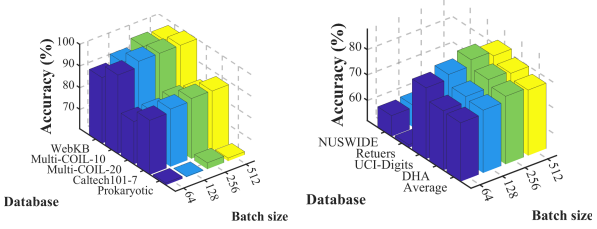
B. Additional Experiment Results

In this section, we list additional results and provides more experimental analysis, which are not shown in the paper due to the space.



(a) (b)

Figure 7: Ablation studies for the different setting of learning rate on the databases. Average denotes the mean accuracy of all the databases.



(a) (b)

Figure 8: Ablation studies for the different setting of batch size on the databases. Average denotes the mean accuracy of all the databases.

More RD Ablation Studies

In this subsection, we provide more ablation experiments for alleviating the RD problem on different databases, as shown in Fig. 6.

We selected the MGBCC (Su et al. 2025) and RMCNC (Sun et al. 2024) for ablation experiments to study the different performance of our proposed method compared with other method on the RD problem. Via observing Fig. 6, we can note that the clustering performance of multi-view fused representation by our method is superior than that of one single view, which not appears in the comparison methods. On the contrary, representation degeneration appears the performance of MGBCC and RMCNC in all the databases.

Ablation of Others Setting

In this subsection, we provide ablation studies for the remaining model parameter settings.

Influence of the learning rate of training. We evaluate the clustering performance of the proposed method with the different learning rates in the process of training, as shown in Fig. 7. We can note that the proposed method achieves the best accuracy in most of all databases when the value is set to $5e-5$.

Influence of the batch size of training. We evaluate the clustering performance of the proposed method with the different batch sizes in the process of training, as shown in Fig. 8. we can observe that the proposed method achieves a nice accuracy performance when the batch size is set to 256. Too small a batch size causes GFCL not being able to focus on enough multi-view samples, which affects the performance of contrastive learning. Increasing the batch size helps with convergence stability, but the generalization performance of the model decreases as the batch size increases.

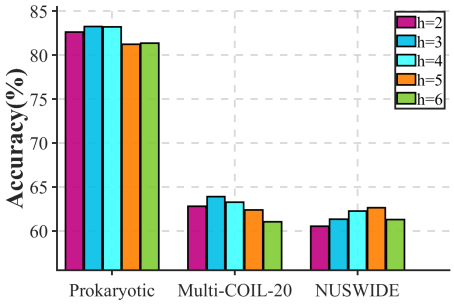
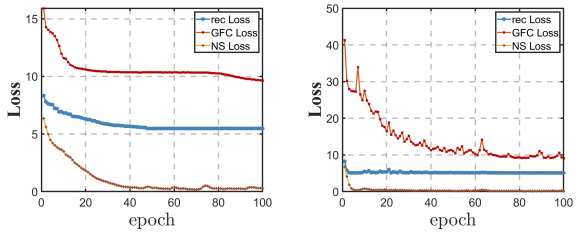


Figure 9: Ablation studies for the different setting of attention heads (h) on the databases.



(a) Influence in WebKB (b) Influence in Caltech101-7

Figure 10: Experiments of convergence analysis on different databases.

Influence of the attention heads of VFS. We evaluate the performance of the proposed method with the different attention heads in the process of VFS, as shown in Fig. 8. We obtain that both excessively high and excessively low h are detrimental to the fusion performance. We set the number of h equal to V , which encourages view-level attention focus.

Convergence Analysis

In this subsection, convergence analysis experiments are performed on the databases used for our experiments. As shown in Fig. 10, the increasing number of epochs leads to a gradual convergence of all loss functions, ultimately reaching a stable state. This clear observation serves as compelling evidence for the stability and effectiveness of our proposed model.