Salient-Residual Decoupled Multi-View Learning for Clustering

Gaokai Wang[1], **Yazhou Ren**[1,2*], **Fengyu Zhang**[1], **Jie Xu**[3], **Chaoning Zhang**[1], **Zhen Long**[4] and **Ce Zhu**[4]

[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China

[2]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China

[3]Information Systems Technology and Design Pillar, Singapore University of Technology and Design

[4]School of Information and Communication Engineering, University of Electronic Science and Technology of China

gaokaiwanggg@gmail.com, yazhou.ren@uestc.edu.cn, 2024190503033@std.uestc.edu.cn, jiexuwork@outlook.com, chaoningzhang1990@gmail.com, {zhen.long, eczhu}@uestc.edu.cn

Abstract

Multi-view clustering aims to utilize information from multiple feature representations to uncover underlying data structures. Most existing methods emphasize learning a consensus representation by enforcing consistency across views. However, those structures that cannot be directly incorporated into the clustering space receive little attention, and are often discarded as noise in practice. In this paper, we argue that such information can be informative signals for the clustering process and should be explicitly modeled rather than suppressed or ignored. Specifically, we propose salient-residual decoupled multi-view learning for clustering, SRDMVC, introducing a novel decomposition-fusion iterative optimization, which separates the feature space into a salient space and a residual subspace effectively and fuses them using a novel attention mechanism. The residual subspace can capture the deep-level structure of view information, making positive contributions to the clustering process, under proper constraints. Without assuming stronger view alignment or complementarity, SRDMVC enhances cluster discrimination, avoids spurious consensus, and alleviates representation degradation. Extensive experiments on benchmark datasets demonstrate the superior performance of the method we propose.

1 Introduction

In the era of high informatization, the lack of label information brings difficulty in directly applying it to the supervised learning process, while the high cost of manual annotation has driven the rapid development of weakly labeled dependency learning methods [Zhang *et al.*, 2025b; Zhang *et al.*, 2025a]. By integrating information from different perspectives, multi-view clustering methods aim to improve clustering robustness and discrimination compared to single-view approaches [Fang *et al.*, 2023; Ren *et al.*, 2025; Chen *et al.*, 2025a]. However, the redundant information brought by multi-view data makes this become the core challenge: how to maximize the effective information among views [Liu *et al.*, 2019; Liu *et al.*, 2020; Zhang *et al.*, 2023]?

Mainstream approaches enhance the process of learning consensus representations by explicitly establishing consistency between different views, to capture shared clustering structures [Yang and Wang, 2018; Ren *et al.*, 2026]. Despite their effectiveness, consensus-driven methods rely on an implicit assumption: representations that cannot be aligned across views or absorbed into the consensus space are either redundant or detrimental to clustering performance [Lyu *et al.*, 2022]. Therefore, such structures are often suppressed or discarded during optimization. However, real-world data is frequently diverse, and the differences brought by which can also reflect the target information [Wang *et al.*, 2016]. Ignoring these differences may lead to spurious consensus, degraded cluster discrimination, and representation degradation [Liang *et al.*, 2019].

Generally speaking, consensus information and complementary information constitute all the features of a view, and it has been proved that correctly integrating is beneficial to the clustering process [Bai *et al.*, 2024; Dong *et al.*, 2023]. Complementary information typically refers to view-specific cues that enrich cluster descriptions, thereby improving the clustering quality [Liu *et al.*, 2024; Fu *et al.*, 2024; Ren *et al.*, 2024]. To ensure the distinction between consensus and complementary information, in practice, the explicitly irrelevant features are often separated and removed as view noise. Intuitively, such noise is assumed to be uninformative for clustering and should not be integrated into the feature space. While, although some of the complementary information does exist in an explicit way to assist clustering, which is the reason why complementary-based methods are functional [Chen *et al.*, 2022], parts of view information hidden in a deeper level is hard to be utilized. When such information is ignored or discarded during optimization, it may lead to irreversible information loss and degraded clustering performance [Li *et al.*, 2019]. We believe that although

*Corresponding author.

some view information that regarded as noise normally may appear to be in conflict with the clustering process, it can, under appropriate constraints, serve as crucial characteristic signals for utilization [Christoudias *et al.*, 2012]. Based on this perspective, incompatible clustering structures should be explicitly modeled rather than implicitly suppressed.

To this end, we propose a new method named SRDMVC, leveraging residual subspaces under consensus constraints for multi-view clustering. We decompose the feature space into two interacting components: (i) the salient space that directly supports clustering, and (ii) the residual subspace that is less relevant to the current clustering process. The residual subspace contains both implicit view-level information and structures that conflict with the clustering objective. Specifically, SRDMVC adopts a novel decomposition-fusion iterative optimization, which first decouples the feature space and then effectively fuses the features using attention mechanisms. We propose salient-residual decoupled multi-view learning and cross-view salient-residual fusion. Salient-residual decoupled multi-view learning mainly achieves decoupling from the complete feature space to the salient feature space and the residual feature subspace. At this stage, separate networks are used to extract salient features and residual features which are initially decomposed through a new decomposition loss, where we design a residual position encoding to make the residual features have view correlation. Then, a novel salient-uniform constraint is proposed to reinforce the role of both features in clustering, and to ensure the positive contributions of residual features to clustering process. Inspired by the attention mechanism in existing clustering approaches [Ma *et al.*, 2023; Zhang and Zhu, 2023; Diallo *et al.*, 2023], cross-view salient-residual fusion uses a salient-residual cross-view attention mechanism to fuse features from different views, and achieves the reconstruction. The proposed method can avoid spurious consensus caused by overemphasizing consistency, enhance clustering discrimination, and alleviate representation degradation.

Our main contributions are as follows:

- We propose salient-residual decoupled multi-view learning, which divides the feature space into salient space and residual subspace, to avoid the information loss that occurred in previous methods when decomposing the common information and complementary information.
- We design cross-view salient-residual fusion, which leverages a salient-residual cross-view attention to fuse features, enabling the effective information in residual subspace to assist the clustering process.
- Extensive experiments on multiple benchmark datasets demonstrate the superior performance of our method, effectively alleviating spurious consensus and representation degradation.

2 Related Work

2.1 Multi-View Clustering

Multi-view clustering (MVC) aims to exploit multiple feature representations of the same data to reveal robust underlying structures. Existing methods can be broadly categorized according to their modeling strategies. Representative works employ low-rank constraints for subspace clustering [Zhang *et al.*, 2015], joint sparse representation with spectral clustering [Khan *et al.*, 2023], or efficient anchor-based graph fusion [Liu *et al.*, 2022]. Scalable and parameter-free fusion frameworks, such as GCFAgg [Yan *et al.*, 2023], have also been widely adopted. A commonality among these methods is their treatment of information inconsistent with the learned consensus as noise to be suppressed. DSMVC [Tang and Liu, 2022] introduces a safety mechanism to mitigate performance degradation caused by increasing views. Contrastive learning, as a powerful method for multi-view clustering, has repeatedly optimized in terms of sample construction and learning approach, and still holds considerable potential [Wang *et al.*, 2025; He *et al.*, 2025]. For the core challenge of extracting consensus information between views in multi-view clustering, methods such as CoMVC [Trosten *et al.*, 2021] and MFLVC [Xu *et al.*, 2022] maximize feature agreement between different views. In order to better utilize the complementary information among views and alleviate the phenomenon of representation degradation, granular learning is introduced into clustering to obtain better feature representations [Su *et al.*, 2025].

2.2 Consistency Constraint

Consistency constraint-based methods explicitly enforce agreement among different views at either the representation, prediction, or clustering assignment level, aiming to learn view-invariant yet discriminative structures. Early works such as Co-regularized Spectral Clustering [Kumar *et al.*, 2011] encourage clustering results from different views to be mutually consistent through shared regularization terms. DCCAE [Wang *et al.*, 2015] extends canonical correlation analysis to deep networks by maximizing cross-view correlation while preserving reconstruction consistency. DMVC [Zhao *et al.*, 2017] introduces a deep framework that aligns latent representations across views using consistency losses to reduce view discrepancy. DAMC [Li *et al.*, 2019] employs adversarial learning to enforce distribution-level consistency between view-specific embeddings and a shared latent space. More recently, the relationship between noise and consistency has received more attention, and more methods attempt to alleviate the noise interference in clustering. Methods such as RMCNC [Sun *et al.*, 2024] multi-view contrastive learning method with noise tolerance, exploring the non-consistency issues in multi-view clustering, while MVCAN [Xu *et al.*, 2024] designs a multi-view iterative optimization method to enhance the robustness of the training process. CANDY [Guo *et al.*, 2024] integrates noise information into the context and utilizes the similarity between different views to discover misclassified samples, alleviating noise interference.

3 Method

3.1 Overview

SRDMVC constructs a decomposition-fusion iterative optimization target, decoupling the feature space into a salient space and a residual subspace. We first decouples the feature

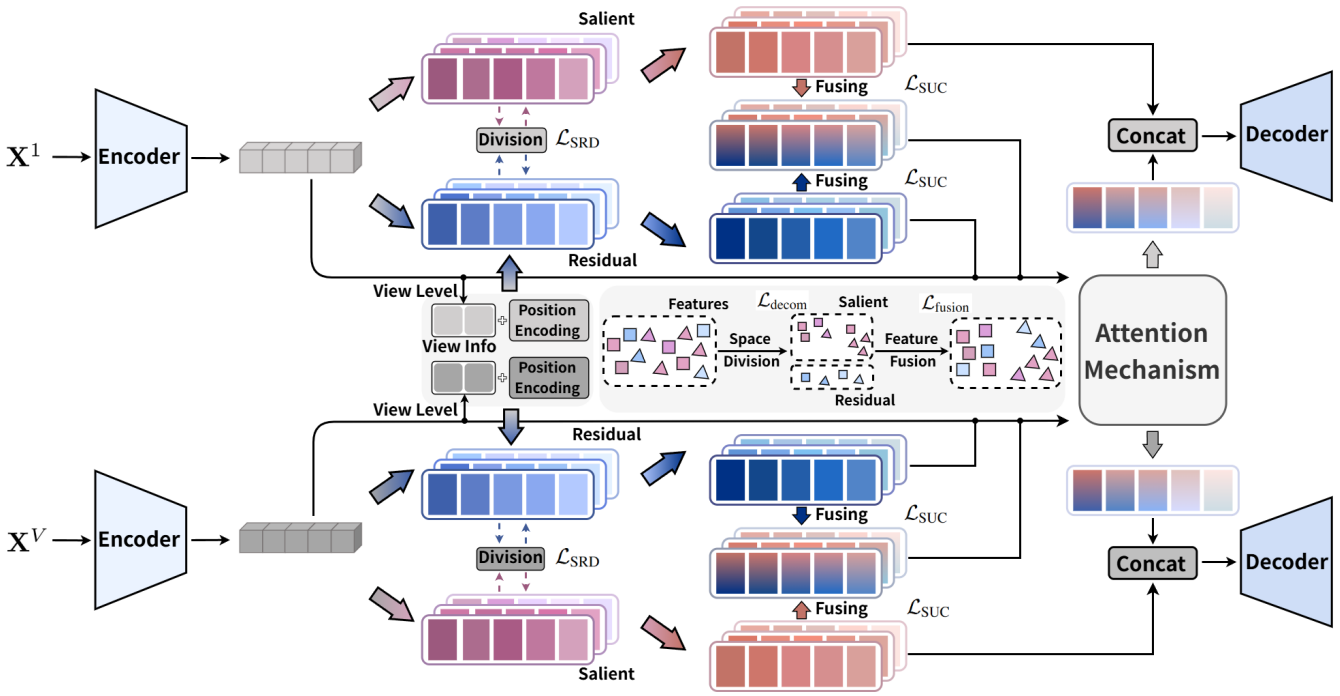


Figure 1: The overview of SRDMVC. SRDMVC divides the feature space into the salient space and the residual subspace, constrains residual features to ensure the effectiveness for clustering process, and leverages an attention mechanism to enhance feature fusion, eventually finishes the reconstruction of views.

space and then effectively fuses the features using attention mechanisms, to ensure the effectiveness of residual features and the interactivity among features. The overall architecture is shown in Figure 1. Given a multi-view dataset $\{\mathbf{X}^i\}_{i=1}^V$ with N samples, and each sample has V different views. During salient-residual decoupled multi-view learning, for each view, samples are processed by a novel network to capture salient and residual features. The we design a correlation-removed loss to decompose different features, and propose a residual position encoding to make the residual features have view correlation. Moreover, we design a salient-uniform constraint to further optimize the correlation between different representations and clustering structures. In cross-view salient-residual fusion, the decoupled features are fused by a salient-residual cross-view attention mechanism, and then fed into decoders to achieve the reconstruction. SRDMVC is trained under decomposition-fusion iterative optimization.

3.2 Salient-Residual Decoupled Multi-View Learning

Model Architecture

For the feature space of a view, common approaches are to divide it into consensus features and complementary features, often accompanied by a simple separation of noise, which may ignore hidden deep information. Our model aims to explicitly decouple each view into a salient feature space and a residual feature subspace to fully extract the information. Specifically, given a multi-view dataset with V views, let $\{\mathbf{X}^i\}_{i=1}^V$ denote the input features of each view. For the i-th view, an independent encoder $f_{\text{enc}}^{(i)}(\cdot)$ is employed to extract initial features, and further obtain unified high-level representations. The process can be expressed as:

$$\mathbf{X}^i \to \mathbf{H}^i \to \mathbf{Z}^i \tag{1}$$

where $\mathbf{H}^i \in \mathbb{R}^{B\times D_H}$, $\mathbf{Z}^i \in \mathbb{R}^{B\times D}$, B represents the batch size of training, and D_H is the feature dimension. During the decomposition process, instead of directly enforcing consensus on $\mathbf{Z}^i$, we transform it into two interacting spaces through parallel transformation modules with different structures. The salient space and residual subspace are extracted through two separate MLPs:

$$\mathbf{S}^i = \text{MLP}_{\text{sal}}^{(i)}(\mathbf{Z}^i), \tag{2}$$

$$\mathbf{R}^i = \text{MLP}_{\text{res}}^{(i)}(\mathbf{Z}^i). \tag{3}$$

$\mathbf{S}^i \in \mathbb{R}^{B\times D}$ is designed to preserve compact and stable structures that can provide direct information for clustering. When considering the fusion of multiple views, consensus information, due to its stability, sharing, and directness, is more likely extracted into salient space. In contrast, $\mathbf{R}^i \in \mathbb{R}^{B\times D}$ contains more structures that cannot be explicitly helpful for clustering, which are hidden within the high-level view information and can only function under specific constraints. Furthermore, we use a linear layer $f_{\text{cat}}^{(i)}(\cdot)$ to fuse the features within the view:

$$\mathbf{F}^i = f_{\text{cat}}^{(i)}([\mathbf{S}^i||\mathbf{R}^i]), \tag{4}$$

where $[\cdot||\cdot]$ denotes feature concatenation, and $\mathbf{F}^i \in \mathbb{R}^{B\times D}$ represents an intermediate embedding, integrating salient and residual information.

Salient-Residual Decomposition

Considering the salient features and residual features obtained from the model, we hope that each feature in $\mathbf{S}^i$ and $\mathbf{R}^i$ should be as complementary as possible, in order to capture more comprehensive information. Furthermore, since the residual features should be weakly associated with the current clustering process, there exists a risk of degeneration into unstructured or view-agnostic representations. Therefore, we set an optimization objective $\mathcal{L}_{\text{SRD}}$ to address the aforementioned issues, which can be defined as:

$$\mathcal{L}_{\text{SRD}} = \sum_{i=1}^{V} \sum_{j=i+1}^{V} \left\| \text{Corr}\left(\mathbf{S}^i, \mathbf{S}^j\right) + \text{Corr}\left(\mathbf{R}^i, \mathbf{R}^j\right) \right\|_F^2 + \sum_{i=1}^{V} \mathcal{R}^r(\mathbf{R}^i). \tag{5}$$

$\text{Corr}(\cdot)$ denotes the covariance calculation, and $\mathcal{R}^r(\cdot)$ represents a regularization method for residual features. Specifically, to ensure that the residual information in each view is encoded into distinct subspaces, we propose the Residual Position Encoding approach, which ensures that the features of each view are only relevant to the current view.

Residual Position Encoding. The main idea is to assign a unique position label to each feature in the residual subspace, based on the specific view it comes from, enabling us to separate residual features from different views. For each view i, the residual features $\mathbf{R}^i$ are encoded with a corresponding position label. Given a multi-view dataset containing V views and the residual feature of the i-th view $\mathbf{R}^i \in \mathbb{R}^{B\times D}$, each residual feature will be assigned a one-hot encoded label $y^i \in \mathbb{R}^{B\times V}$ as the label of view i, which can be defined as:

$$y^i = [\mathbf{0}, ..., \underbrace{\mathbf{1}}_{i-\text{th}}, ..., \mathbf{0}], \tag{6}$$

where $\mathbf{1}$ is at the i-th position and $\mathbf{0}$ elsewhere, $\mathbf{0}$ and $\mathbf{1}$ are vectors of length B. The position labels are assigned based on the index of each view, such that each feature in a given view has a view-unique label. Residual features $\mathbf{R}^i$ are passed through a residual view discriminator $f_{dis}(\cdot)$ to classify the features according to their view. The process can be written as:

$$\hat{y}^i = f_{\text{dis}}(\mathbf{R}^i). \tag{7}$$

$\hat{y}^i \in \mathbb{R}^{B\times V}$ represents the discriminator's output logits of view i. To optimize the target, we use the cross-entropy loss as the regularization method to ensure that the residual features are correctly classified according to their corresponding view, which can be expressed as:

$$\mathcal{R}^r(\mathbf{R}^i) = \frac{1}{BV} \sum_{i=1}^{V} \sum_{j=1}^{B} \text{CrossEntropy}(\hat{y}^i, y^i), \tag{8}$$

where $\hat{y}^i$ is the predicted label and y^i is the true one-hot label corresponding to its view. Overall, the residual subspace is encouraged to retain the distinguishable features across views by minimizing $\mathcal{L}_{\text{SRD}}$, thereby preventing it from degenerating into single or invalid representations. It is important to note that this enhancement mechanism does not guarantee that residual features are complementary. Instead, it ensures structural effectiveness and view relevance, which plays a crucial role in the following stage.

Salient-Uniform Constraint

In this stage, we do not enforce alignment between residual features and clustering assignments, but aim to prevent the residual features from distorting clustering process. Specifically, we introduce the Salient-Uniform Constraint to achieve three primary goals: (i) The salient features should directly reflect the clustering structure; (ii) The residual features should be uncorrelated with the clustering process; (iii) The fused features should have stronger clustering capabilities. And to establish a unified clustering reference, we introduce a shared learnable clustering head $g(\cdot)$, which maps inputs to a soft cluster assignment. For each view i, the clustering head is applied to salient features $\mathbf{S}^i$, residual features $\mathbf{R}^i$, and fused features $\mathbf{F}^i$ to make consistent evaluation across spaces. The process can be expressed as:

$$\mathbf{Q}_s^i = g(\mathbf{S}^i), \mathbf{Q}_r^i = g(\mathbf{R}^i), \mathbf{Q}_f^i = g(\mathbf{F}^i). \tag{9}$$

Among them, $\mathbf{Q}_s^i$ and $\mathbf{Q}_f^i$ can directly reflect the clustering structures, while $\mathbf{Q}_r^i$ should maintain a low correlation with the clustering structure. In other words, the salient and fused features are expected to be highly aligned with the clustering assignments, while the residual features are designed to be less directly associated with clustering structures. In practice, we propose a loss function to achieve these targets, which is formalized by the following Salient-Uniform Constraint loss:

$$\mathcal{L}_{\text{SUC}}^i = D_{KL}(\mathbf{Q}_s^i || \mathbf{P}_s^i) - \sum_{k=1}^{K} \bar{q}_k^i \log(\bar{q}_k^i) + \sum_{i=1}^{V} \mathcal{R}^s(\mathbf{S}^i), \tag{10}$$

where $D_{KL}(\cdot)$ represents the KL divergence, $\mathbf{P}_s^i$ is the squared normalization of $\mathbf{Q}_s^i$, $\mathcal{R}^s(\cdot)$ denotes a regularization method for salient features $\mathbf{S}^i$, and $\bar{\mathbf{q}}^i = \frac{1}{B}\sum_{j=1}^{B} \mathbf{Q}_{r,j}^i$. The first term $D_{KL}(\mathbf{Q}_s^i || \mathbf{P}_s^i)$ ensures that the structures can be helpful, the second term $-\sum_{k=1}^{K} \bar{q}_k^i \log(\bar{q}_k^i)$ penalizes the residual features for deviating from a uniform distribution, which does not allow for residual features to have clustering preferences, the third term $\sum_{i=1}^{V} \mathcal{R}^s(\mathbf{S}^i)$ forces the strongest clustering ability for fused features.

Salient Clustering Regularization. To guarantee that the interaction between salient and residual subspaces leads to positive results, we require the clustering structures reflected by $\mathbf{F}^i$ should not be inferior to that of the salient representation alone, which means that residual features should make positive contributions to clustering targets after initial feature fusion. So we design a regularization method for salient features:

$$\mathcal{R}^s(\mathbf{S}^i) = \max\left(0, D_{KL}(\mathbf{Q}_s^i || \mathbf{P}_s^i) - D_{KL}(\mathbf{Q}_f^i || \mathbf{P}_f^i)\right), \tag{11}$$

where $\mathbf{P}_f^i$ is the squared normalization of $\mathbf{Q}_f^i$.

3.3 Cross-View Salient-Residual Fusion

Salient-Residual Cross-View Attention

In previous steps, the feature space of the view was decoupled and optimized through decomposition loss. Furthermore, in order to better facilitate the interaction and mutual promotion among multiple views, we first propose salient-residual cross-view attention to construct a unified and view-aware representation. Traditional fusion approaches may lose intrinsic connections between different features, which is not conducive capturing the relationship between salient space and residual subspace. Inspired by the multi-view attention mechanism, we utilize a variant to adaptively aggregate features. We consider three types of representations: $\mathbf{H}$, $\mathbf{R}$, and $\mathbf{F}$, which are organized into a query–key–value triplet to facilitate structured information aggregation. These features are respectively obtained by concatenating all the corresponding view features $\mathbf{H}^i$, $\mathbf{R}^i$, and $\mathbf{F}^i$. Specifically, $\mathbf{H}$ encodes the semantic context, $\mathbf{R}$ captures residual structures regulated by consensus constraints, and $\mathbf{F}$ provides the candidate representations to be fused. Finally, these representations are input into the attention mechanism to obtain fused features, which can be formulated as:

$$\mathrm{Attn}(\mathbf{H}, \mathbf{R}, \mathbf{F}) = \mathrm{softmax}(\frac{(\mathbf{H})^\top \times \mathbf{R}}{\sqrt{D}})\mathbf{F}, \tag{12}$$

where attention is performed independently for each head and then aggregated through concatenation and linear projection. Unlike conventional multi-view fusion strategies, which rely on fixed weighting or direct concatenation, salient-residual cross-view attention explicitly conditions the aggregation process on residual cues and view-semantic representations, while allowing the fusion process to selectively emphasize views whose residual structures are informative under the current representation context, which provides a robust foundation for reconstruction.

View-Wise Reconstruction

To keep the effectiveness of the integration across views, we employ a view-wise reconstruction loss. The reconstruction process requires the utilization of $\mathbf{F}^i$, which provides regulatory information for fusion. While during the attention fusing process, the salient feature may be weakened due to the characteristics of attention and the limitations of input structures. To alleviate this, the fused feature $\mathbf{X}_{att} = \mathrm{Attn}(\mathbf{H}, \mathbf{R}, \mathbf{F})$ is combined with the salient feature $\mathbf{S}^i$ to obtain the final fused representation:

$$\mathbf{F}^i_{\text{fin}} = [\mathbf{X}^i_{\text{att}} || \mathbf{S}^i]. \tag{13}$$

$\mathbf{X}^i_{\text{att}}$ is the attention-fused feature of view i from $\mathbf{X}_{\text{att}}$. Subsequently, we use a view-independent decoder $f^{(i)}_{dec}(\cdot)$ to decode the fusion feature:

$$\hat{\mathbf{X}}^i = f^{(i)}_{\text{dec}}(\mathbf{F}^i_{\text{fin}}), \tag{14}$$

where $\hat{\mathbf{X}}^i$ denotes the decoded data of view v. By employing independent decoders, the model avoids forcing a single one to fit heterogeneous feature distributions, which maintains flexibility across views. For the formulation of reconstruction loss, we leverage the Mean Squared Error (MSE) to accomplish, which can be defined as:

Dataset	Size	View	Class
WebKB	2102	2-views	2-classes
Prokaryotic	551	3-views	4-classes
MSRC-v5	210	5-views	7-classes
ACM	3025	5-views	3-classes
Wikipedia	2866	2-views	10-classes
Multi-COIL-20	1440	3-views	20-classes

Table 1: Details of the benchmark datasets used in experiments

$$\mathcal{L}^v_{\text{REC}} = \| \mathbf{X}^i - \hat{\mathbf{X}}^i \|_2^2 . \tag{15}$$

Decomposition-Fusion Iterative Optimization

In SRDMVC, salient-residual decoupled multi-view learning and cross-view salient-residual fusion are two independent iterative processes, trained to achieve the decomposition-fusion iterative optimization. During the Salient-Residual Decoupled Multi-View Learning stage, we first decouple the salient and residual features, ensuring that they are learned independently for each view. Once this stage is completed, the learned weights are inherited and continue the training process in the Cross-View Salient-Residual Fusion stage, where the features from multiple views are fused for clustering. Overall, the decomposition-fusion iterative optimization can be expressed as:

$$\begin{cases} \mathcal{L}_{\text{decom}} & = \mathcal{L}_{\text{SRD}} + \sum_{i=1}^{V} \mathcal{L}^i_{\text{SUC}}, \\ \mathcal{L}_{\text{fusion}} & = \sum_{i=1}^{V} \mathcal{L}^i_{\text{REC}}. \end{cases} \tag{16}$$

During the actual training process, we utilize some hyperparameters to adjust the weights of different terms in $\mathcal{L}^i_{\text{SUC}}$ to optimize the learning objective. Based on this, $\mathcal{L}^i_{\text{SUC}}$ can be written as:

$$\mathcal{L}^i_{\text{SUC}} = D_{KL}(\mathbf{Q}^i_s || \mathbf{P}^i_s) - \alpha \sum_{k=1}^{K} \bar{q}^i_k \log(\bar{q}^i_k) + \beta \sum_{i=1}^{V} \mathcal{R}^s(\mathbf{S}^i). \tag{17}$$

α and β are hyperparameters used to adjust the regulating intensity. By minimizing $\mathcal{L}_{\text{decom}}$, SRDMVC can decouple the salient space and residual subspace while maintaining positive contributions of residual features for clustering. And the optimization of $\mathcal{L}_{\text{fusion}}$ makes the cross-view feature fusion more effective.

4 Experiments

4.1 Experimental Setup

Datasets

The datasets involved in our experiments are shown in Table 1. In particular, WebKB [Sun *et al.*, 2007] is a multi-view web page dataset, where each instance is extracted from various textual and structural sources. Prokaryotic [Brbić *et al.*, 2016] describes prokaryotic organisms using multiple genomic feature representations, aiming to categorize samples based on functional or taxonomic similarities. MSRC-v5 [Xia *et al.*, 2023] is a widely used image segmentation

Method	WebKB				Prokaryotic				MSRC-v5			
	ACC	NMI	ARI	PUR	ACC	NMI	ARI	PUR	ACC	NMI	ARI	PUR
DSMVC (2022)	71.08	24.54	17.71	78.12	58.62	22.65	22.47	68.24	47.65	38.75	26.28	51.90
SEM (2023)	60.70	1.88	3.73	78.11	49.73	22.22	11.90	58.80	51.43	47.64	30.53	52.86
GCFAgg (2023)	59.18	3.63	3.30	59.18	57.71	30.09	21.38	57.71	36.49	31.92	15.60	40.48
SCMVC (2024)	95.81	69.65	82.39	95.81	58.80	23.47	26.06	67.51	45.71	39.18	23.74	49.05
DFL-NET (2025)	77.83	36.33	30.63	78.12	48.64	20.92	11.91	62.79	39.05	39.40	25.75	40.00
MS2FA-CMVC (2025)	71.46	29.59	18.36	78.12	48.28	25.58	17.29	67.33	14.29	0	0	14.29
GFSAF-AE (2026)	**98.29**	83.28	88.16	**98.29**	63.89	35.49	32.15	76.59	44.76	37.93	21.62	48.10
GFSAF-DAE (2026)	94.13	70.24	80.06	94.13	62.07	33.18	31.72	**84.03**	54.29	43.49	29.63	56.67
SRDMVC (ours)	98.19	**83.98**	**90.98**	98.19	**77.50**	**49.36**	**49.13**	**84.03**	**67.62**	**53.59**	**42.85**	**67.62**

Method	ACM				Wikipedia				Multi-COIL-20			
	ACC	NMI	ARI	PUR	ACC	NMI	ARI	PUR	ACC	NMI	ARI	PUR
DSMVC (2022)	59.04	18.56	20.31	59.04	59.17	53.83	45.69	62.93	81.18	87.06	75.51	81.25
SEM (2023)	64.83	41.09	34.27	64.83	25.54	9.37	5.60	27.00	81.18	86.36	76.09	83.82
GCFAgg (2023)	65.34	41.02	35.84	65.34	51.29	41.03	31.47	51.29	82.80	90.18	80.32	82.80
SCMVC (2024)	66.91	44.43	38.64	66.91	56.94	55.63	46.20	60.89	93.89	**98.39**	93.98	95.00
DFL-NET (2025)	68.43	41.09	41.18	68.43	27.46	20.01	11.12	29.10	29.72	67.77	32.02	32.43
MS2FA-CMVC (2025)	59.50	25.58	24.71	59.50	45.46	34.00	26.60	52.09	82.48	88.16	80.19	82.96
GFSAF-AE (2026)	60.24	30.16	25.64	60.24	58.41	54.56	45.11	62.98	85.49	89.29	79.83	85.69
GFSAF-DAE (2026)	56.40	24.93	24.10	56.89	59.80	**56.40**	42.77	64.65	80.63	86.20	74.73	80.76
SRDMVC (ours)	**68.83**	**47.91**	**42.29**	**68.83**	**61.27**	55.76	**46.29**	**65.60**	**97.85**	98.12	**96.05**	**97.85**

Table 2: Performance comparison of different methods on public multi-view datasets. The best and second best results are indicated by bold and underline, respectively.

dataset where each image is described by several visual features. ACM [Jin *et al.*, 2021] is a citation network dataset, in which the documents are represented from different relational and content-based perspectives. Wikipedia [Rasiwasia *et al.*, 2010] is a large multi-view document dataset constructed from Wikipedia pages, where each article is described by multiple sets of different modalities of features. Multi-COIL-20 [Nene *et al.*, 1996] are multi-view object image datasets with each photo of an object taken from multiple perspectives is regarded as a different view.

Baselines

To verify the performance of proposed SRDMVC, we compare it with eight different multi-view methods for clustering, including: DSMVC [Tang and Liu, 2022], SEM [Xu *et al.*, 2023], GCFAgg [Yan *et al.*, 2023], SCMVC [Wu *et al.*, 2024], DFL-NET [Chen *et al.*, 2025b], MS2FA-CMVC[Zhang *et al.*, 2025c], and two different backbones of GFSAF (GFSAF-AE, GFSAF-DAE).

Evaluation Metrics

We evaluate the clustering performance of all comparison methods by adopting four widely-used metrics: Clustering Accuracy (ACC), Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), and Purity (PUR).

Implementation Details

In terms of training parameters, we set the learning rate to 1e-4, and the batch size for all datasets to 256, and the Adam optimizer is used to train model. In Eq.(17), α and β are respectively set to 0.1 and 0.8. For feature dimension, the value of D is set to 64, and the value of D_H is set to $V \times D$. All experiments are conducted on a workstation equipped with an NVIDIA RTX 2080 Ti GPU (11 GB memory).

Loss Strategies			WebKB			Prokaryotic		
$\mathcal{L}'_{\text{SRD}}$	$\mathcal{L}_{\text{SUC}}$	$\mathcal{R}^r$	ACC	NMI	PUR	ACC	NMI	PUR
✓	-	-	96.57	74.37	96.57	66.97	37.93	76.59
-	-	✓	95.14	72.68	95.14	58.08	32.64	68.06
-	✓	✓	97.72	80.26	97.72	70.78	45.83	83.85
✓	-	✓	97.24	76.98	97.24	71.51	39.17	78.95
✓	✓	-	96.19	71.64	96.19	73.87	45.27	82.03
✓	✓	✓	98.19	83.98	98.19	77.50	49.36	84.03

Table 3: Ablation results of SRDMVC's loss strategies on dataset WebKB and Prokaryotic.

4.2 Comparison

Table 2 shows the performance comparison between SRDMVC and baseline methods. Our method achieves the optimal or near-optimal results on all datasets. Overall, SRDMVC has average improvements of on ACC, NMI, ARI, and PUR, respectively. On view-heterogeneous and weakly aligned datasets, which are more challenging, SRDMVC also achieves outstanding performance, which demonstrates the effectiveness of the modeling architecture we perform.

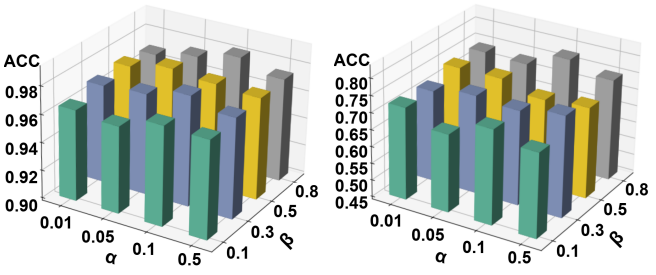
4.3 Ablation Study

Influence of Loss Modules

To verify the effectiveness of the loss components of SRDMVC, we conduct experiments using different combinations of loss terms, which are shown in Table 3. $\mathcal{L}'_{\text{SRD}}$ represents

Residual Form	WebKB			Prokaryotic		
	ACC	NMI	PUR	ACC	NMI	PUR
w/o Residual	96.57	74.94	96.57	71.32	47.04	83.48
Random Residual	93.63	58.86	93.63	65.34	37.44	77.56
Gaussian Residual	94.48	62.33	94.48	70.78	46.66	83.30
Full Model	98.19	83.98	98.19	77.50	49.36	84.03

Table 4: Ablation results of residual subspace in SRDMVC on dataset WebKB and Prokaryotic.



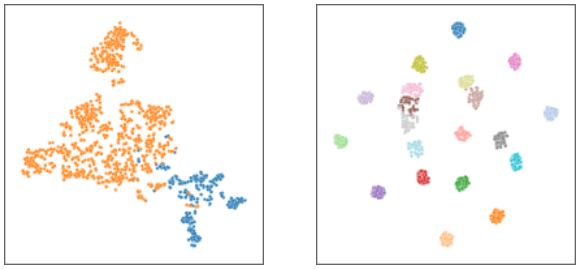
(a) WebKB (b) Prokaryotic

Figure 2: Influence of different combinations of α and β for clustering performance on dataset WebKB and Prokaryotic.

the $\mathcal{L}_{\text{SRD}}$ without the regulation method $\mathcal{R}^r$. When $\mathcal{L}'_{\text{SRD}}$ is employed, the result is the highest when only one loss module is activated. Experiments with combined loss terms reveal that introducing $\mathcal{L}_{\text{SUC}}$ consistently leads to a substantial performance improvement, and further introduction of $\mathcal{R}^r$ provides additional gains for $\mathcal{L}_{\text{SRD}}$. These observations indicate that explicitly introducing the residual subspace into the view fusion process makes positive contributions to improving clustering performance. Finally, fusing all loss functions achieves the overall best performance of SRDMVC.

Effectiveness of Residual Subspace

To investigate the contribution of residual subspace in SRDMVC, we conducted a series of ablation experiments by removing or modifying the residual representation while keeping other components unchanged, as shown in Table 4. The residual subspace is defined as the feature sets that cannot explicitly contribute to the clustering process, which might be discarded as noises in traditional approaches. Therefore, in practice, we consider the following four variants: (1) w/o Residual, where the residual subspace is completely removed and only salient space is retained; (2) Random Residual, where the residual representation is initialized with randomly sampled noise following a uniform distribution, aiming to verify whether performance improves stem merely from increased feature dimensionality; (3) Gaussian Residual, where the residual subspace is replaced by Gaussian noise of zero mean and unit variance; (4) Full Model, where the residual subspace is explicitly learned and jointly optimized with salient space. The comparison between removing the residual subspace and using the full model demonstrates the positive contribution of residual features.



(a) WebKB (b) Multi-COIL-20

Figure 3: Visualization of the fused feature on WebKB and Multi-COIL-20.

Influence of Loss Weights

The value of $\mathcal{L}_{\text{SUC}}$ is affected by α and β. In this section, we used different combinations of alpha and beta to investigate the impact of loss weights on the model performance, as shown in Figure 2. Specifically, we evaluate four values for α, i.e., $\{0.01, 0.05, 0.1, 0.5\}$, and four values for β, i.e., $\{0.1, 0.3, 0.5, 0.8\}$, on WebKB and Prokaryotic. Experimental results show that when α is set too small, the salient features of the residual subspace might not be completely "erased". Conversely, overly large weights can also degrade performance, which may indicate that the residual subspace collapses toward uninformative noise. $\mathcal{R}^s$ encourages that the residual subspace should be conducive to feature fusion. As β increases, this effect will be enhanced, allowing the residual subspace to make more contributions to feature fusion. Overall, we select $\alpha = 0.1$ and $\beta = 0.8$ as the best combination, which achieves the optimal performance of SRDMVC.

Visualization of Final Fusion Features

We visualize the fused features to show the clustering result, as illustrated in Figure 3, which shows the competitive clustering results.

5 Conclusion

In this paper, we propose a novel framework for multi-view clustering, called SRDMVC. Compared with traditional approaches, SRDMVC introduces a salient-residual decoupled method to enhance feature fusion, which alleviates the information loss that occurs during the feature decomposition process when using traditional methods. In particular, we adopt a novel decomposition-fusion iterative optimization. We first decouple the feature space to salient space and residual subspace, leverage proper constraints to reinforce the positive contributions of both features in clustering, and ensure the effectiveness of residual features. Further, we fuse the decoupled features through a salient-residual cross-view attention mechanism, and achieve the effective reconstruction for clustering structures. As shown in the experiments, SRDMVC shows outstanding performance, demonstrating the effectiveness of our method. Future work will continue to focus on the distribution of residual features, and develop more reasonable optimization.

Acknowledgments

This work was supported in part by National Key Research and Development Program of China (No. 2024YFC2310801) and National Natural Science Foundation of China (No. 62476052).

References

[Bai *et al.*, 2024] Ruina Bai, Ruizhang Huang, Yongbin Qin, Yanping Chen, and Yong Xu. A structural consensus representation learning framework for multi-view clustering. *KNOWL-BASED SYST*, 283:111132, 2024.

[Brbić *et al.*, 2016] Maria Brbić, Matija Piškorec, Vedrana Vidulin, Anita Kriško, Tomislav Šmuc, and Fran Supek. The landscape of microbial phenotypic traits and associated genes. *NUCLEIC ACIDS RES*, page gkw964, 2016.

[Chen *et al.*, 2022] Man-Sheng Chen, Chang-Dong Wang, Dong Huang, Jian-Huang Lai, and Philip S Yu. Efficient orthogonal multi-view subspace clustering. In *KDD*, pages 127–135, 2022.

[Chen *et al.*, 2025a] Jianpeng Chen, Yawen Ling, Jie Xu, Yazhou Ren, Shudong Huang, Xiaorong Pu, Zhifeng Hao, Philip S Yu, and Lifang He. Variational graph generator for multiview graph clustering. *IEEE T NEUR NET LEAR*, 36(6):11078–11091, 2025.

[Chen *et al.*, 2025b] Zhe Chen, Xiao-Jun Wu, Tianyang Xu, and Josef Kittler. Dfl-net: Disentangled feature learning network for multi-view clustering. *TKDE*, 2025.

[Christoudias *et al.*, 2012] C Christoudias, Raquel Urtasun, and Trevor Darrell. Multi-view learning in the presence of view disagreement. *arXiv preprint arXiv:1206.3242*, 2012.

[Diallo *et al.*, 2023] Bassoma Diallo, Jie Hu, Tianrui Li, Ghufran Ahmad Khan, Xinyan Liang, and Hongjun Wang. Auto-attention mechanism for multi-view deep embedding clustering. *PATTERN RECOGN*, 143:109764, 2023.

[Dong *et al.*, 2023] Zhibin Dong, Siwei Wang, Jiaqi Jin, Xinwang Liu, and En Zhu. Cross-view topology based consistent and complementary information for deep multi-view clustering. In *ICCV*, pages 19440–19451, 2023.

[Fang *et al.*, 2023] Uno Fang, Man Li, Jianxin Li, Longxiang Gao, Tao Jia, and Yanchun Zhang. A comprehensive survey on multi-view clustering. *TKDE*, 35(12):12350–12368, 2023.

[Fu *et al.*, 2024] Lele Fu, Sheng Huang, Lei Zhang, Jinghua Yang, Zibin Zheng, Chuanfu Zhang, and Chuan Chen. Subspace-contrastive multi-view clustering. *ACM T KNOWL DISCOV D*, 18(9):1–35, 2024.

[Guo *et al.*, 2024] Ruiming Guo, Mouxing Yang, Yijie Lin, Xi Peng, and Peng Hu. Robust contrastive multi-view clustering against dual noisy correspondence. *NeurIPS*, 37:121401–121421, 2024.

[He *et al.*, 2025] Hongqing He, Jie Xu, Guoqiu Wen, Yazhou Ren, Na Zhao, and Xiaofeng Zhu. Graph embedded contrastive learning for multi-view clustering. In *IJCAI*, pages 5336–5344, 2025.

[Jin *et al.*, 2021] Di Jin, Cuiying Huo, Chundong Liang, and Liang Yang. Heterogeneous graph neural network via attribute completion. In *WWW*, pages 391–400, 2021.

[Khan *et al.*, 2023] Ghufran Ahmad Khan, Jie Hu, Tianrui Li, Bassoma Diallo, and Shengdong Du. Multi-view subspace clustering for learning joint representation via low-rank sparse representation. *APPL INTELL*, 53(19):22511–22530, 2023.

[Kumar *et al.*, 2011] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. *NeurIPS*, 24, 2011.

[Li *et al.*, 2019] Zhaoyang Li, Qianqian Wang, Zhiqiang Tao, Quanxue Gao, Zhaohua Yang, et al. Deep adversarial multi-view clustering network. In *IJCAI*, volume 2, page 4, 2019.

[Liang *et al.*, 2019] Youwei Liang, Dong Huang, and Chang-Dong Wang. Consistency meets inconsistency: A unified graph learning framework for multi-view clustering. In *ICDM*, pages 1204–1209, 2019.

[Liu *et al.*, 2019] Xinwang Liu, Xinzhong Zhu, Miaomiao Li, Chang Tang, En Zhu, Jianping Yin, and Wen Gao. Efficient and effective incomplete multi-view clustering. In *AAAI*, volume 33, pages 4392–4399, 2019.

[Liu *et al.*, 2020] Xinwang Liu, Miaomiao Li, Chang Tang, Jingyuan Xia, Jian Xiong, Li Liu, Marius Kloft, and En Zhu. Efficient and effective regularized incomplete multi-view clustering. *TPAMI*, 43(8):2634–2646, 2020.

[Liu *et al.*, 2022] Suyuan Liu, Siwei Wang, Pei Zhang, Kai Xu, Xinwang Liu, Changwang Zhang, and Feng Gao. Efficient one-pass multi-view subspace clustering with consensus anchors. In *AAAI*, volume 36, pages 7576–7584, 2022.

[Liu *et al.*, 2024] Maoshan Liu, Vasile Palade, and Zhonglong Zheng. Learning the consensus and complementary information for large-scale multi-view clustering. *NEURAL NETWORKS*, 172:106103, 2024.

[Lyu *et al.*, 2022] Gengyu Lyu, Xiang Deng, Yanan Wu, and Songhe Feng. Beyond shared subspace: A view-specific fusion for multi-view multi-label learning. In *AAAI*, volume 36, pages 7647–7654, 2022.

[Ma *et al.*, 2023] Zhifeng Ma, Junyang Yu, Longge Wang, Huazhu Chen, Yuxi Zhao, Xin He, Yingqi Wang, and Yalin Song. Multi-view clustering based on view-attention driven. *INT J MACH LEARN CYB*, 14(8):2621–2631, 2023.

[Nene *et al.*, 1996] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-100). Technical report, 1996.

[Rasiwasia *et al.*, 2010] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM MM*, pages 251–260, 2010.

[Ren *et al.*, 2024] Yazhou Ren, Xinyue Chen, Jie Xu, Jingyu Pu, Yonghao Huang, Xiaorong Pu, Ce Zhu, Xiaofeng Zhu,

Zhifeng Hao, and Lifang He. A novel federated multi-view clustering method for unaligned and incomplete data fusion. *INFORM FUSION*, 108:102357, 2024.

[Ren *et al.*, 2025] Yazhou Ren, JunLong Ke, Zichen Wen, Tianyi Wu, Yang Yang, Xiaorong Pu, and Lifang He. Multi-view graph clustering via node-guided contrastive encoding. In *ICML*, 2025.

[Ren *et al.*, 2026] Yazhou Ren, Zichen Wen, Junlong Ke, Chenhang Cui, Yonghao Huang, Xinyue Chen, Philip S Yu, and Lifang He. Enhancing multi-view clustering: A sufficient information-theoretic approach for consistency acquisition and redundancy elimination. *TPAMI*, 2026.

[Su *et al.*, 2025] Peng Su, Shudong Huang, Weihong Ma, Deng Xiong, and Jiancheng Lv. Multi-view granular-ball contrastive clustering. In *AAAI*, volume 39, pages 20637–20645, 2025.

[Sun *et al.*, 2007] Ting-Kai Sun, Song-Can Chen, Zhong Jin, and Jing-Yu Yang. Kernelized discriminative canonical correlation analysis. In *ICWAPR*, volume 3, pages 1283–1287, 2007.

[Sun *et al.*, 2024] Yuan Sun, Yang Qin, Yongxiang Li, Dezhong Peng, Xi Peng, and Peng Hu. Robust multi-view clustering with noisy correspondence. *TKDE*, 2024.

[Tang and Liu, 2022] Huayi Tang and Yong Liu. Deep safe multi-view clustering: Reducing the risk of clustering performance degradation caused by view increase. In *CVPR*, pages 202–211, 2022.

[Trosten *et al.*, 2021] Daniel J Trosten, Sigurd Lokse, Robert Jenssen, and Michael Kampffmeyer. Reconsidering representation alignment for multi-view clustering. In *CVPR*, pages 1255–1265, 2021.

[Wang *et al.*, 2015] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *ICML*, pages 1083–1092. PMLR, 2015.

[Wang *et al.*, 2016] Hao Wang, Yan Yang, and Tianrui Li. Multi-view clustering via concept factorization with local manifold regularization. In *ICDM*, pages 1245–1250. IEEE, 2016.

[Wang *et al.*, 2025] Qianqian Wang, Haiming Xu, Zihao Zhang, Zhiqiang Tao, and Quanxue Gao. Efficient multi-view clustering via reinforcement contrastive learning. In *IJCAI*, pages 6361–6369, 2025.

[Wu *et al.*, 2024] Song Wu, Yan Zheng, Yazhou Ren, Jing He, Xiaorong Pu, Shudong Huang, Zhifeng Hao, and Lifang He. Self-weighted contrastive fusion for deep multi-view clustering. *TMM*, 26:9150–9162, 2024.

[Xia *et al.*, 2023] Wei Xia, Quanxue Gao, Qianqian Wang, Xinbo Gao, Chris Ding, and Dacheng Tao. Tensorized bipartite graph learning for multi-view clustering. *TPAMI*, 45(4):5187–5202, 2023.

[Xu *et al.*, 2022] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *CVPR*, pages 16051–16060, 2022.

[Xu *et al.*, 2023] Jie Xu, Shuo Chen, Yazhou Ren, Xiaoshuang Shi, Hengtao Shen, Gang Niu, and Xiaofeng Zhu. Self-weighted contrastive learning among multiple views for mitigating representation degeneration. *NeurIPS*, 36:1119–1131, 2023.

[Xu *et al.*, 2024] Jie Xu, Yazhou Ren, Xiaolong Wang, Lei Feng, Zheng Zhang, Gang Niu, and Xiaofeng Zhu. Investigating and mitigating the side effects of noisy views for self-supervised clustering algorithms in practical multi-view scenarios. In *CVPR*, pages 22957–22966, 2024.

[Yan *et al.*, 2023] Weiqing Yan, Yuanyang Zhang, Chenlei Lv, Chang Tang, Guanghui Yue, Liang Liao, and Weisi Lin. Gcfagg: Global and cross-view feature aggregation for multi-view clustering. In *CVPR*, pages 19863–19872, 2023.

[Yang and Wang, 2018] Yan Yang and Hao Wang. Multi-view clustering: A survey. *Big Data Min Anal*, 1(2):83–107, 2018.

[Zhang and Zhu, 2023] Yanhao Zhang and Changming Zhu. Incomplete multi-view clustering via attention-based contrast learning. *INT J MACH LEARN CYB*, 14(12):4101–4117, 2023.

[Zhang *et al.*, 2015] Changqing Zhang, Huazhu Fu, Si Liu, Guangcan Liu, and Xiaochun Cao. Low-rank tensor constrained multiview subspace clustering. In *ICCV*, pages 1582–1590, 2015.

[Zhang *et al.*, 2023] Lei Zhang, Lele Fu, Tong Wang, Chuan Chen, and Chuanfu Zhang. Mutual information-driven multi-view clustering. In *CIKM*, pages 3268–3277, 2023.

[Zhang *et al.*, 2025a] Malu Zhang, Shuai Wang, Jibin Wu, Wenjie Wei, Dehao Zhang, Zijian Zhou, Siying Wang, Fan Zhang, and Yang Yang. Toward energy-efficient spike-based deep reinforcement learning with temporal coding. *IEEE COMPUT INTELL M*, 20(2):45–57, 2025.

[Zhang *et al.*, 2025b] Malu Zhang, Wenjie Wei, Zijian Zhou, Wanlong Liu, Jie Zhang, Ammar Belatreche, and Yang Yang. Spike-driven lightweight large language model with evolutionary computation. *IEEE T EVOLUT COMPUT*, 2025.

[Zhang *et al.*, 2025c] Yuanyang Zhang, Weiqing Yan, Chang Tang, Wujie Zhou, and Jian Jin. Multi-branch space sharing feature aggregation for contrastive multi-view clustering. *PATTERN RECOGN*, page 111704, 2025.

[Zhao *et al.*, 2017] Handong Zhao, Zhengming Ding, and Yun Fu. Multi-view clustering via deep matrix factorization. In *AAAI*, volume 31, 2017.