RELATIONSHIP ALIGNMENT FOR VIEW-AWARE MULTI-VIEW CLUSTERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view clustering improves clustering performance by integrating complementary information from multiple views. However, existing methods often suffer from two limitations: i) the neglect of preserving sample neighborhood structures, which weakens the consistency of inter-sample relationships across views; and ii) inability to adaptively utilize inter-view similarity, resulting in representation conflicts and semantic degradation. To address these issues, we propose a novel framework named Relationship Alignment for View-aware Multi-view Clustering (RAV). Our approach first constructs a sample relation matrix for each view using deep features and aligns it with a global relation matrix to enhance neighborhood consistency across views. Furthermore, we introduce a view-aware adaptive weighting mechanism for label contrastive learning. This mechanism dynamically adjusts the contrastive intensity between view pairs based on the similarity of their deep features: higher similarity leads to stronger label alignment, while lower similarity reduces the weighting to prevent forcing inconsistent views into agreement. This strategy effectively promotes cluster-level semantic consistency while preserving natural inter-view relationships. Extensive experiments demonstrate that our method consistently outperforms state-of-the-art approaches on multiple benchmark datasets.

1 INTRODUCTION

In recent years, with the rapid development of big data and multi-modal data, multi-view clustering (MVC) Chen et al. (2023b); Dong et al. (2023); Eisenberg et al. (2025); Trosten et al. (2023); Wan et al. (2024) has emerged as a significant research direction and has been widely applied across various domains, including computer vision Xie et al. (2020), natural language processing Ke et al. (2024); Nadkarni et al. (2011), and social network analysis Fang et al. (2023b); Banez et al. (2022). Unlike traditional clustering methods that rely on single data representations, MVC methods can more comprehensively capture intrinsic data relationships and latent structures by effectively integrating complementary information from different views, thereby achieving more accurate sample partitioning. Existing MVC approaches can be broadly categorized into two types: traditional MVC methods and deep MVC methods.

With the continuous advancement of deep learning across various domains, deep MVC Xu et al. (2021); Trosten et al. (2021); Chen et al. (2025a); Xiao et al. (2025) has gradually become mainstream. Compared to traditional MVC, deep MVC leverages the powerful representation learning capabilities of deep neural networks and has demonstrated promising clustering performance. For instance, Lin et al. (2021) learns view-specific representations through intra-view reconstruction loss while maintaining cross-view consistency using mutual information-based contrastive learning. Xu et al. (2022b) proposes a multi-level feature learning framework that alleviates the conflict between learning consistent representations and reconstructing inconsistent features by learning low-level features, high-level features, and semantic labels. Yan et al. (2023) achieves consistent representation learning through inter-sample structural relationships and employs similarity structures to guide contrastive learning. These approaches effectively capture complex nonlinear relationships across views, thereby significantly improving both feature representation quality and clustering efficiency.

Meanwhile, contrastive learning Chen et al. (2025c; 2020; 2024; 2025b) plays a significant role in representation learning. Current contrastive learning-based deep MVC approaches primarily

focus on two levels: sample-level and cluster-level. At the sample level, these methods enhance feature consistency by pulling similar samples closer and pushing dissimilar ones apart. At the cluster level, they aim to achieve consistency in multi-view cluster distributions. For example, Li et al. (2021) introduces both sample-level and cluster-level contrastive objectives to jointly optimize feature representations and clustering assignments. Chen et al. (2023a) proposes cross-view cluster assignment contrastive learning to achieve consistent cluster distributions. Cui et al. (2024) develops a dual contrastive mechanism that systematically constructs positive and negative pairs for learning consistent representations. These research contributions provide novel insights for multi-view clustering tasks, while establishing a solid foundation for subsequent contrastive learning-based deep MVC methodologies.

Despite considerable advances in deep multi-view clustering (MVC), several challenges remain. Many existing methods do not adequately preserve sample neighborhood structures across views. Furthermore, when views exhibit substantial discrepancies, naively applying contrastive learning can cause representation conflicts and distort semantic information, ultimately degrading clustering performance. Although recent works Xu et al. (2023); Wu et al. (2024) have introduced adaptive view weighting for feature-level contrastive learning, such approaches often neglect the alignment of inter-view sample relations and fail to harness view similarity to guide label-level contrastive learning—without disrupting inherent view relationships. These limitations hinder the learning of consistent and discriminative representations. To overcome these issues, we propose a novel framework that combines sample relationship alignment with view-aware adaptive label contrastive learning. Our method begins by extracting deep features from each view using view-specific encoders. We then construct a relation matrix per view—as well as a global relation matrix—using a Gaussian kernel to capture sample affinities. By aligning each view's relation matrix with the global matrix, we enforce cross-view consistency while preserving local neighborhood structures. Additionally, cluster assignment matrices are generated via a shared MLP. A view-aware weighting strategy is introduced to modulate the strength of label contrastive learning between view pairs based on their feature similarity. Pairs with higher similarity receive greater emphasis to strengthen label consistency, while those with lower similarity are down-weighted to avoid harmful forcing. This approach effectively reduces representation degradation caused by view discrepancies and maintains the natural relationships between views.

The main contributions of this paper are summarized as follows:

- We introduce a global-guide-local sample relation alignment module that preserves neighborhood structures and enhances cross-view consistency by aligning view-specific relation matrices with a global relation matrix.
- We propose a view-aware adaptive weighting mechanism for label contrastive learning, which dynamically emphasizes high-similarity view pairs to strengthen semantic consistency, while reducing the influence of low-similarity pairs to prevent representation degradation.
- Extensive experiments on multiple benchmarks show that our method achieves state-of-the-art performance across standard clustering metrics, demonstrating its effectiveness and generalizability.

2 Related Work

2.1 Deep Multi-View Clustering

The advancement of deep learning has greatly propelled the development of deep MVC. Modern deep MVC methods capitalize on the strong nonlinear representation ability of neural networks to learn shared latent representations from multi-view data effectively. These approaches can be broadly categorized into several lines of research. Graph-based methods, such as the graph structure-aware contrastive clustering proposed by Fei et al. (2025), enhance representation learning by incorporating sample-level topological relationships alongside attribute features. Subspace-based methods focus on deriving consistent representations across views; for example, Yu et al. (2025) introduces a pseudo-label-guided bidirectional discriminative subspace clustering framework that uses pseudo-label-driven contrastive learning and a dual-attention mechanism to maintain structural coherence in sample affinity matrices. Reconstruction-based methods learn representations by recovering view-specific features or structures. Xu et al. (2022a); Yan et al. (2025), for instance, applies a

self-supervised strategy that reconstructs view features to generate pseudo-labels, which in turn guide the learning of discriminative multi-view features through a unified target distribution.

2.2 CONTRASTIVE LEARNING

Contrastive learning has emerged as a powerful unsupervised paradigm for handling data heterogeneity. By constructing positive and negative sample pairs, it learns discriminative and consistent representations, showing particular strength in cross-view alignment tasks. Dong et al. (2025) presents a view-graph-based progressive fusion method with dual contrastive learning within and across views, enabling consistent multi-view representation learning. Cui et al. (2024) designs a dual contrastive loss that combines a dynamic clustering diffusion term to separate clusters and a neighbor-guided alignment term to improve within-cluster compactness. Tang et al. (2020) further proposes a decoupled contrastive MVC approach based on higher-order graph walks, which learns reliable cluster representations through concurrent intra-view and inter-view contrastive learning.

Despite these advances, existing methods often overlook the alignment of sample-level relational structures and do not adaptively utilize view similarity to guide label-aware contrastive learning. As a result, they struggle to preserve neighborhood consistency across views and are susceptible to representation degradation when view discrepancies are large. In contrast, our approach explicitly aligns the relational structures of individual views with a global relation matrix to maintain cross-view neighborhood consistency. Moreover, we introduce a view-aware adaptive weighting strategy that modulates the contribution of view pairs in label contrastive learning based on their feature similarity, thereby enhancing semantic consistency without distorting inherent view relationships.

3 METHOD

3.1 PRELIMINARIES

This section introduces our multi-view clustering framework, as depicted in Figure 1. The architecture consists of three core components: View-Specific Autoencoder Modules, a Cross-View Relation Alignment Module, and a View-aware Label Contrastive Learning Module. Given a multi-view dataset with V views denoted as $\mathbf{X} = \{\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^V\}$, where the data from the v-th view is represented as $\mathbf{X}^v = [\mathbf{x}_1^v; \mathbf{x}_2^v; \ldots; \mathbf{x}_N^v] \in \mathbb{R}^{N \times d_v}$, with N being the number of samples and d_v the feature dimension of view v, the framework is designed to jointly optimize these modules. This integrated approach effectively handles disparities in view similarity while promoting consistency in cross-view representations.

3.2 VIEW-SPECIFIC AUTOENCODER

In multi-view clustering, the quality of feature representations is critical to clustering performance. To extract robust latent features from raw multi-view data, which often contain noise and redundancy, we employ view-specific autoencoders for each view. Formally, for the v-th view, we define an encoder f^v and a decoder g^v. The latent representation of the i-th sample in view v is obtained as::

$$\mathbf{z}_i^v = f^v(\mathbf{x}_i^v; \theta^v), \tag{1}$$

where θ^v denotes the learnable parameters of the encoder for the v-th view, and d is the dimensionality of the resulting latent feature. The latent representation $\mathbf{z}_i^v$ is then passed to the corresponding decoder g^v to reconstruct the original input, formulated as:

$$\hat{\mathbf{x}}_i^v = g^v(\mathbf{z}_i^v; \phi^v) = g^v(f^v(\mathbf{x}_i^v; \theta^v); \phi^v), \tag{2}$$

where $\hat{\mathbf{x}}_i^v$ denotes the reconstructed version of the i-th sample from the v-th view, and ϕ^v represents the trainable parameters of its decoder. The overall reconstruction loss for training all autoencoders is defined as:

$$\mathcal{L}_{REC} = \sum_{v=1}^{V} \sum_{i=1}^{N} \|\mathbf{x}_i^v - \hat{\mathbf{x}}_i^v\|_2^2. \tag{3}$$

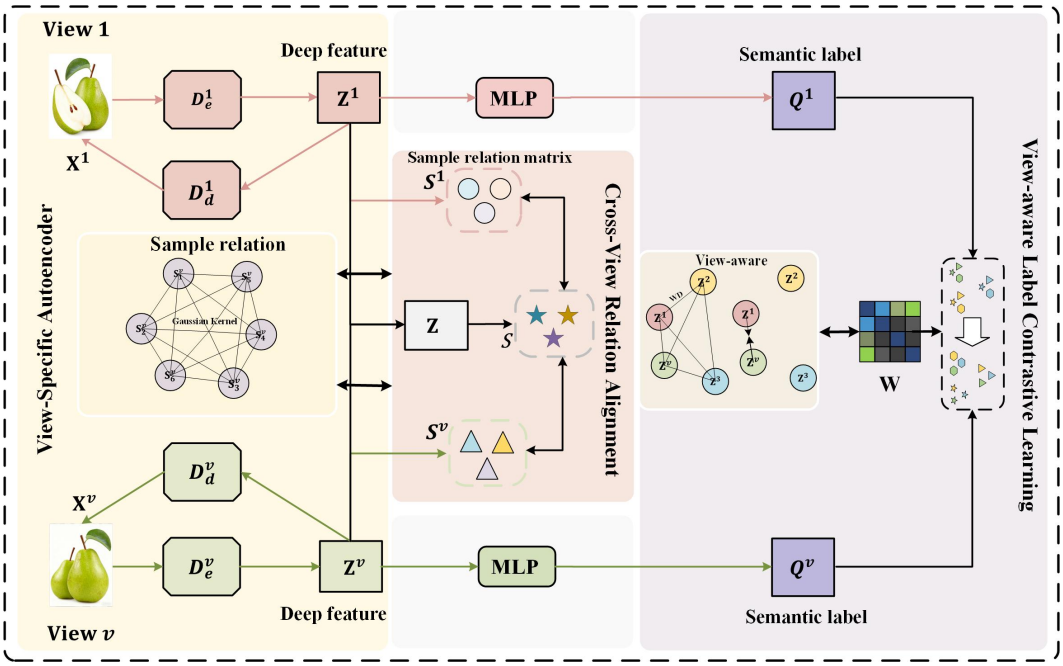


Figure 1: An illustration of the proposed RAV framework. The model crucially incorporates two modules: cross-view relation alignment to maintain neighborhood structures, and view-aware adaptive weighting in label contrastive learning to counteract representation degradation from view dissimilarity, thereby achieving robust multi-view clustering.

3.3 Cross-view Relationship Alignment

To enhance the consistency of sample relationships across views while preserving local neighborhood structures, we introduce a cross-view relation alignment module. First, deep features $\mathbf{Z}^v$ for each view are obtained using the view-specific encoders. The pairwise similarity between samples within a view is then computed via a Gaussian kernel. Specifically, the similarity between the i-th and k-th samples in the v-th view is given by:

$$s_{ik}^v = \exp\left(-\frac{\|\mathbf{z}_i^v - \mathbf{z}_k^v\|^2}{\sigma}\right), \tag{4}$$

where s_{ik}^v denotes the similarity between the i-th and k-th samples in the v-th view. A smaller feature distance corresponds to higher similarity, and vice versa. Based on equation (4), we compute the pairwise similarities for each view and store them in a view-specific relation matrix $\{\mathbf{S}^v = [\mathbf{s}_1^v; \mathbf{s}_2^v; , , , ; \mathbf{s}_N^v]\}_{v=1}^V \in \mathbb{R}^{N\times N}$. To construct a global relation matrix that integrates information from all views, we first concatenate the deep features as follows:

$$\mathbf{Z} = \text{Concat}(\mathbf{Z}^1, \mathbf{Z}^2, \ldots, \mathbf{Z}^v), \tag{5}$$

where $\mathbf{Z} \in \mathbb{R}^{N\times(Vd)}$ represents the concatenated global features. The global relation matrix $\mathbf{S} = [\mathbf{s}_1; \mathbf{s}_2; , , , ; \mathbf{s}_N] \in \mathbb{R}^{N\times N}$ is then computed using the same similarity measure defined in equation (4). We align each view-specific relation matrix with this global matrix via a global-supervise-local contrastive learning objective, which pulls positive pairs closer while pushing negative pairs apart. The resulting cross-view relation alignment loss is formulated as:

$$\mathcal{L}_{\text{S}} = -\frac{1}{N}\sum_{v=1}^{V}\sum_{i=1}^{N}\log\frac{e^{d(\mathbf{s}_i^v, \mathbf{s}_i)/\tau_F}}{\sum_{k=1}^{N} e^{d(\mathbf{s}_i^v, \mathbf{s}_k)/\tau_F} - e^{1/\tau_L}}, \tag{6}$$

where τ_F is a temperature hyperparameter, and $d(\mathbf{s}_i^v, \mathbf{s}_k)$ denotes the cosine similarity function, defined as:

$$d(\mathbf{s}_i^v, \mathbf{s}_k) = \frac{\langle \mathbf{s}_i^v, \mathbf{s}_k \rangle}{\|\mathbf{s}_i^v\| \|\mathbf{s}_k\|}. \tag{7}$$

The alignment loss improves cross-view relational consistency by preserving sample neighborhood structures, ensuring that neighboring samples remain close while distant ones are separated. Consequently, it enhances both feature discriminability within views and semantic consistency across views.

3.4 View-aware Label Contrastive Learning

After obtaining the view-specific representations, we project the features of each view through a shared MLP to generate the cluster assignment matrices $\{\mathbf{Q}^v \in \mathbb{R}^{N \times K}\}_{v=1}^V$. Each entry $\mathbf{q}_{ij}^v$ denotes the probability that the i-th sample is assigned to the j-th cluster in the v-th view, obtained by applying the Softmax function along the cluster dimension. To promote clustering consistency across views, we apply contrastive learning at the cluster-assignment level. For the j-th cluster assignment vector $\mathbf{q}_{:,j}^v$ (i.e., the j-th column of $\mathbf{Q}^v$), we consider all possible assignment pairs across views and clusters, totaling $(VK-1)$ pairs. Among these, pairs originating from the same cluster index j but different views ($u \neq v$) are treated as positive pairs, amounting to $(V-1)$ positives. The remaining $V(K-1)$ pairs are considered negatives. The contrastive loss between view v and view u is defined as:

$$\ell_c^{(v,u)} = -\frac{1}{K} \sum_{j=1}^{K} \log \frac{e^{d(\mathbf{q}_{:,j}^v, \mathbf{q}_{:,j}^u)/\tau_L}}{\sum_{k=1}^{K} \sum_{m=v,u} e^{d(\mathbf{q}_{:,j}^v, \mathbf{q}_{:,k}^m)/\tau_L} - e^{1/\tau_L}}, \tag{8}$$

where τ_L is the temperature parameter of label contrastive learning. Afterward, the total label contrastive loss is give by:

$$\mathcal{L}_{\mathrm{Q}} = \frac{1}{2} \sum_{v=1}^{V} \sum_{u \neq v} \ell_c^{(v,u)} + \sum_{v=1}^{V} \sum_{j=1}^{K} \mathbf{r}_j^v \log \mathbf{r}_j^v, \tag{9}$$

where $\mathbf{r}_j^v = \frac{1}{N} \sum_{i=1}^{N} \mathbf{q}_{ij}^v$. The first term enhances cross-view clustering consistency, while the second acts as a regularization term that prevents all samples from being assigned to a single cluster, thereby avoiding trivial solutions.

However, this approach does not account for inherent feature distribution discrepancies across views. When contrastive learning forcibly aligns view pairs with substantial differences, it may distort genuine semantic structures and cause representation degradation. To address this issue, we propose a view-aware adaptive weighting strategy for label contrastive learning, which dynamically modulates the influence of each view pair based on their deep feature similarity. Specifically, we first employ the Wasserstein Distance (WD) Shen et al. (2018) to quantify the discrepancy between the feature distributions of two views. The WD between view v and view u is defined as:

$$WD(\mathbf{Z}^v, \mathbf{Z}^u) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{k=1}^{N} |\mathbf{z}_i^v - \mathbf{z}_k^u|, \tag{10}$$

where $\mathbf{z}_i^v$ and $\mathbf{z}_k^u$ denote the deep features of the i-th sample in view v and the k-th sample in view u, respectively. The adaptive weight between views v and u is then calculated based on their Wasserstein Distance as follows:

$$w_{(v,u)} = \frac{e^{-WD(\mathbf{Z}^v, \mathbf{Z}^u)}}{\sum_{v=1}^{V} e^{-WD(\mathbf{Z}^v, \mathbf{Z}^u)}}, \tag{11}$$

where $w_{(v,u)}$ denotes the adaptive weight between the representations $\mathbf{Z}^v$ and $\mathbf{Z}^u$. All pairwise weights are organized into a $V \times V$ matrix $\mathbf{W}$. This matrix enables a dynamic weighting strategy: view pairs with high feature similarity (i.e., small WD values) are assigned larger weights to strengthen their contribution during contrastive learning, while pairs with large representation discrepancies (i.e., high WD values) are assigned smaller weights to reduce potential negative effects. Integrating this

weighting mechanism into the label contrastive learning objective, we obtain the final view-aware adaptive weighting loss:

$$\mathcal{L}_{\mathrm{Q}} = \frac{1}{2}\sum_{v=1}^{V}\sum_{u\neq v}\frac{1}{2}(w_{(v,u)} + w_{(u,v)})\ell_c^{(v,u)} + \sum_{v=1}^{V}\sum_{j=1}^{K}\mathbf{r}_j^v \log \mathbf{r}_j^v. \tag{12}$$

3.5 The Overall Loss Function

Based on the foregoing formulation, the overall objective function integrates the three key components as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{REC}} + \lambda_1\mathcal{L}_{\mathrm{Q}} + \lambda_2\mathcal{L}_{\mathrm{S}}, \tag{13}$$

where λ_1 and λ_2 are balancing coefficients, $\mathcal{L}_{\text{REC}}$ denotes the reconstruction loss, $\mathcal{L}_{\mathrm{Q}}$ represents the view-aware adaptive label contrastive loss, and $\mathcal{L}_{\mathrm{S}}$ corresponds to the cross-view relation alignment loss.

Once the model converges, the clustering labels can be obtained as follows:

$$y_j = \arg\max_j \left(\frac{1}{V}\sum_{v=1}^{V}\mathbf{q}_{i,j}^v\right). \tag{14}$$

The full process of our RAV is summarized in Algorithm 1.

Algorithm 1 : The optimization of RAV.

1: **Input**: Multi-view dataset $\{\mathbf{X}^v\}_{v=1}^V$; The number of samples is N; The number of max epochs is T; The number of clusters is K; The parameters λ_1, λ_2.
2: **Initialization:** Initialize autoencoder parameters by minimizing $\mathcal{L}_{\text{REC}}$ in Eq. (3).
3: **for** $t = 1$ to T **do**
4: Obtain the weight matrix $\mathbf{W}$ by Eq. (11).
5: Optimize $\{\theta^v, \phi^v\}_{v=1}^V$ by minimizing $\mathcal{L}_{\text{total}}$ in Eq. (13) .
6: **end for**
7: Calculate the predicted labels by Eq. (14).
8: **Output**: $\mathbf{Y} = [y_1, y_2, \ldots, y_N]$.

4 Experiment

4.1 Datasets and Experimental Setting

Datasets. We evaluate our model on nine benchmark datasets. Table 1 summarizes their key characteristics, including sample size, number of views, number of clusters, and feature dimensions for each view. NGs Yan et al. (2025): This dataset consists of 500 documents, which have been preprocessed using three different methods to obtain three distinct views. Digit-Product Xu et al. (2021): This dataset is derived from MNIST and Fashion Handwritten digits, containing 30,000 samples and two views. ALOI Cui et al. (2024): This dataset contains 10,800 samples and 10 clusters, with four views extracted from each image, representing color similarity, Haralick, HSV, and RGB features. Cora Fang et al. (2023a): This dataset contains 2,708 documents, with four features selected as the four views: content, inbound, outbound, and citations. It is categorized into seven clusters. NUSWIDE Chua et al. (2009): This dataset consists of 5,000 images, classified into 5 categories. Caltech-5V Xu et al. (2022b): This dataset is an RGB image dataset containing 1,400 images, covering WM, CENTRIST, LBP, GIST, and HOG features. NoisyMNIST Wang et al. (2015): This dataset comprises 50,000 samples, organized into 10 clusters. YoutubeVideo Madani et al. (2012): This dataset consists of 101,499 samples, divided into 31 classes. 3Sources[1]: This dataset contains 169 samples, 3 views, and 6 classes.

[1] http://mlg.ucd.ie/datasets/3sources.html

Table 1: Description of the used multi-view datasets.

Dataset	Samples	Views	Clusters	Dimensionality
NGs	500	3	5	2000/2000/2000
Digit-Product	30,000	2	10	1024/1024
ALOI	10,800	4	100	77/13/64/125
Cora	2,708	4	7	2708/1433/2708/2708
NUSWIDE	5,000	5	5	65/226/145/74/129
Caltech-5V	1,400	5	7	40/254/928/512/1984
NoisyMNIST	50,000	2	10	784/784
YoutubeVideo	101,499	3	31	512/647/838
3Sources	169	3	6	3560/3631/3068

Implementation Details. All experiments are conducted using PyTorch 1.12.1 on an NVIDIA RTX 4090 D GPU. The model is optimized with the Adam optimizer, employing a fixed learning rate of 0.0003 and a batch size of 256. Both pretraining and fine-tuning phases are fixed at 200 epochs. We introduce two hyperparameters: λ_1 and λ_2, where λ_1 ranges from $[0.00001, 0.0001, \ldots, 1000]$ and λ_2 ranges from $[0.00001, 0.0001, \ldots, 1]$. All baseline methods were evaluated under identical experimental conditions.

4.2 Compared Methods and Results

We compare seven representative multi-view clustering methods across nine benchmark datasets to evaluate our approach. MFLVC Xu et al. (2022b): This method is primarily used for multi-level feature learning in multi-view clustering. GCFAgg Yan et al. (2023): This method mainly utilizes sample similarity structures to guide contrastive learning. SEM Xu et al. (2023): This method mostly guides contrastive learning by adjusting view weights. MVCAN Xu et al. (2024): This method alleviates the negative impact of noisy views and optimizes the learning of individual image representations. DDMVC Xu et al. (2025): This method considers diversity and discriminative feature learning. SSLNMVC Yan et al. (2025): This method introduces the UProjection module, which enhances the expressiveness of consistent features by feature resampling and concatenating the fused features before and after resampling. AICN-MLM Shu et al. (2025): This method proposes a fuzzy instance-aware multi-level matching contrastive network for multi-view document clustering.

We evaluate our method using three widely recognized clustering evaluation metrics: Accuracy (ACC), Normalized Mutual Information (NMI), and Purity (PUR), as shown in Tables 2, 3, and 4. Based on these results, we can draw the following conclusions:

Our method achieves superior performance over baseline approaches on most datasets, confirming its effectiveness. Notably, on the NGs, YoutubeVideo, and 3Sources datasets, it improves ACC by 4.4%, 7.8%, and 1.2%, respectively, over the second-best results. These gains stem from the proposed view-aware adaptive contrastive learning and relation alignment mechanism, which mitigates representation conflicts from view discrepancies, emphasizes high-similarity view pairs in label learning, and maintains neighborhood consistency—collectively enhancing clustering accuracy.

On ALOI and Caltech-5V, our method performs slightly below MVCAN, which may be due to the latter's non-use of standard contrastive learning, reducing its sensitivity to view variations. On the Fashion dataset, our results are comparable to MFLVC, SEM, AICN-MLM, and SSLNMVC, likely because of the dataset's limited inter-view variability, diminishing the need for adaptive weighting.

Compared to SEM, which also uses view similarity in feature contrastive learning, our method performs better on all datasets except Fashion. This highlights the stronger generalizability of our deep feature similarity mechanism in guiding label contrastive learning. Our approach not only enhances the contribution of high-similarity views but also preserves natural inter-view relationships, leading to improved robustness and generalization.

4.3 Model Analyses

Parameter Sensitivity: To evaluate the impact of λ_1 and λ_2 on the model, we conduct a parameter sensitivity analysis as illustrated in Figure 2. The experiments set

Table 2: Clustering results of all methods on the NGs, Digit-Product, and ALOI datasets.

Datasets	NGs			Digit-Product			ALOI		
Evaluation Metrics	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
MFLVC (22 CVPR)	0.932	0.825	0.932	0.991	0.976	0.991	0.435	0.786	0.435
GCFAgg (23 CVPR)	0.894	0.742	0.894	0.988	0.968	0.988	0.790	0.917	0.809
SEM (24 NeurIPS)	0.856	0.673	0.856	0.991	0.976	0.991	0.771	0.899	0.787
MVCAN (24 CVPR)	0.470	0.271	0.470	0.989	0.967	0.989	**0.849**	**0.929**	**0.864**
DDMVC (25 PR)	–	–	–	0.968	0.931	0.968	0.796	0.907	0.813
SSLNMVC (25 TMM)	0.936	0.842	0.936	0.990	0.973	0.990	0.541	0.814	0.558
AICN-MLM (25 AAAI)	0.912	0.774	0.912	0.991	0.976	0.991	0.788	0.906	0.800
ours	**0.980**	**0.934**	**0.980**	**0.998**	**0.993**	**0.998**	0.826	0.912	0.830

Table 3: Clustering results of all methods on the Cora, NUSWIDE, and Caltech-5V datasets.

Datasets	Cora			NUSWIDE			Caltech-5V		
Evaluation metrics	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
MFLVC (22 CVPR)	0.268	0.111	0.377	0.624	0.338	0.624	0.867	0.781	0.867
GCFAgg (23 CVPR)	0.220	0.051	0.304	0.596	0.336	0.596	0.799	0.697	0.799
SEM (24 NeurIPS)	0.220	0.028	0.313	0.588	0.317	0.588	0.901	0.834	0.901
MVCAN (24 CVPR)	0.567	0.385	**0.640**	0.572	0.290	0.572	**0.919**	**0.856**	**0.919**
DDMVC (25 PR)	0.323	0.141	0.407	0.607	0.312	0.636	0.771	0.695	0.779
SSLNMVC (25 TMM)	0.277	0.102	0.364	0.637	0.367	0.637	0.881	0.789	0.881
AICN-MLM (25 AAAI)	0.331	0.171	0.418	0.612	0.337	0.612	0.898	0.828	0.898
ours	**0.592**	**0.404**	0.598	**0.647**	**0.371**	**0.647**	0.901	0.839	0.901

Table 4: Clustering results of all methods on the NoisyMNIST, YoutubeVideo, 3Sources, and Fashion datasets.

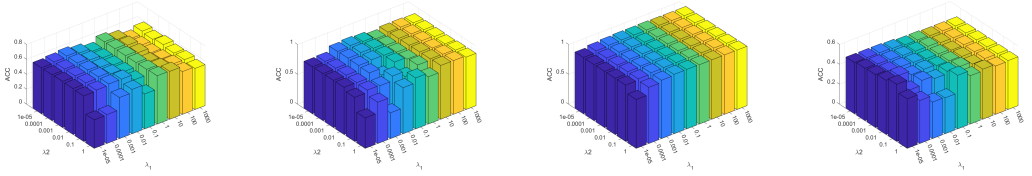
Datasets	NoisyMNIST			YoutubeVideo			3Sources			Fashion		
Evaluation metrics	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
MFLVC (22 CVPR)	0.988	0.965	0.988	0.238	0.224	0.324	0.521	0.477	0.669	**0.994**	**0.985**	**0.994**
GCFAgg (23 CVPR)	0.781	0.847	0.836	0.275	0.263	0.361	0.521	0.429	0.615	0.990	0.974	0.990
SEM (24 NeurIPS)	0.995	0.984	0.995	0.318	0.309	0.404	0.533	0.584	0.716	**0.994**	0.983	**0.994**
MVCAN (24 CVPR)	0.933	0.861	0.933	0.244	0.244	0.341	0.562	0.478	0.663	0.856	0.840	0.856
DDMVC (25 PR)	0.957	0.893	0.957	–	–	–	0.456	0.346	0.592	0.931	0.892	0.931
SSLNMVC (25 TMM)	0.995	0.985	0.995	0.235	0.244	0.430	0.521	0.510	0.686	**0.994**	0.984	**0.994**
AICN-MLM (25 AAAI)	0.990	0.971	0.990	–	–	–	0.538	0.472	0.675	**0.994**	**0.985**	**0.994**
ours	**0.996**	**0.986**	**0.996**	**0.356**	**0.332**	**0.445**	**0.574**	**0.599**	**0.775**	**0.994**	0.984	**0.994**

Table 5: Ablation studies on different loss components on the Caltech-5V, NUSWIDE, ALOI, and 3Sources datasets.

Components			Caltech-5V			NUSWIDE			ALOI			3Sources		
$\mathcal{L}_{\text{REC}}$	$\mathcal{L}_{\text{Q}}$	$\mathcal{L}_{\text{S}}$	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
✓	✓	×	0.899	0.830	0.899	0.644	0.362	0.644	0.780	0.887	0.789	0.562	0.464	0.686
✓	×	✓	0.424	0.309	0.439	0.298	0.037	0.311	0.264	0.656	0.264	0.396	0.135	0.408
✓	✓	✓	**0.901**	**0.839**	**0.901**	**0.647**	**0.371**	**0.647**	**0.826**	**0.912**	**0.830**	**0.574**	**0.599**	**0.775**

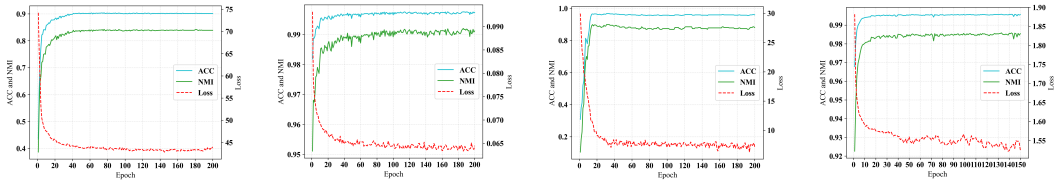
λ_1 within the range $[10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3]$ and λ_2 within the range $[10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1]$. The results demonstrate that when these hyperparameters vary within the specified ranges, the clustering performance on all four datasets exhibits only minor fluctuations. This relatively small variation in performance proves that our method is highly robust to hyperparameter selection.

Convergence: We observe the changes in training loss and evaluation metrics (ACC/NMI) across four benchmark datasets to analyze the convergence characteristics of the proposed method. Figure 3 shows the convergence curves for the Fashion, Hdigit, NUSWIDE, and YTF-10 datasets. The main



(a) NUSWIDE (b) Caltech-5V (c) NoisyMNIST (d) 3Sources

Figure 2: Parameter sensitivity analysis of λ_1 and λ_2 on the NUSWIDE, Caltech-5V, NoisyMNIST, and 3Sources datasets.



(a) Caltech-5V (b) Digit-Product (c) NGs (d) NoisyMNIST

Figure 3: Convergence analysis of ours on Caltech-5V, Digit-Product, NGs, and NoisyMNIST datasets, where each subgraph displays the total loss and both evaluation metrics (ACC/NMI) over training epochs.

observations are as follows: First, the loss function decreases rapidly in the initial training phase and then gradually stabilizes until convergence. Second, the clustering evaluation metrics ACC and NMI continuously increase during training and eventually stabilize. Finally, this convergence trend indicates that our loss function effectively regularizes the model and drives parameter optimization. These results not only demonstrate the convergence stability of our method but also validate its effectiveness in improving clustering performance.

Table 6: Ablation study on the view-aware adaptive weighting mechanism for NGs, Digit-Product, ALOI and Cora datasets.

Datasets	NGs			Digit-Product			ALOI			Cora		
Evaluation metrics	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
ours w/o W	0.966	0.895	0.966	**0.998**	**0.993**	**0.998**	0.801	0.903	0.807	0.585	0.393	0.585
ours	**0.980**	**0.934**	**0.980**	**0.998**	**0.993**	**0.998**	**0.826**	**0.912**	**0.830**	**0.592**	**0.404**	**0.598**

Ablation Studies: To systematically evaluate the contribution of each component, we conduct a comprehensive ablation study. As summarized in Table 5, $\mathcal{L}_{\text{REC}}$, $\mathcal{L}_{\text{Q}}$, and $\mathcal{L}_{\text{S}}$ represent the reconstruction loss, the view-aware adaptive weighting contrastive loss, and the cross-view relation alignment loss, respectively. Experiments on four benchmark datasets—Caltech-5V, NUSWIDE, ALOI, and 3Sources—show that the full model outperforms variants that remove either the relation alignment module or the view-aware weighting mechanism. These results confirm that relation alignment helps capture semantic structures across views, while adaptive weighting enhances robustness by dynamically moderating the influence of view pairs.

We further examine the specific contribution of the view-aware weighting strategy in Table 6. On the NGs, ALOI, and Cora datasets, the full model achieves ACC gains of 1.4%, 2.5%, and 0.7%, respectively, compared to the variant without weighting (Ours w/o W). These improvements indicate that the weighting strategy effectively alleviates representation degradation resulting from view discrepancy. On the Digit-Product dataset, however, performance remains unchanged—likely due to its inherently small inter-view differences, which diminish the need for adaptive weighting. This outcome underscores the particular usefulness of our weighting mechanism in scenarios with pronounced view disparities.

5 CONCLUSION

This paper presents a multi-view clustering framework that integrates sample relation alignment with view-aware adaptive weighting for contrastive learning, leading to significant performance improvements. The framework begins by constructing relation matrices for each view and a global matrix using a Gaussian kernel. It then enforces cross-view consistency and preserves neighborhood structures by aligning each view-specific relation matrix with the global one. Moreover, a view-aware adaptive weighting mechanism based on Wasserstein Distance is introduced to reduce the negative effects of view similarity discrepancies. Extensive experiments show that the proposed method effectively mitigates the impact of view disparity and outperforms existing approaches on multiple benchmarks. In future work, we will explore more generalizable view similarity measurements to develop a universal evaluation standard adaptable to diverse data characteristics, further enhancing the method's applicability and stability in real-world scenarios.

REFERENCES

Reginald A Banez, Hao Gao, Lixin Li, Chungang Yang, Zhu Han, and H Vincent Poor. Modeling and analysis of opinion dynamics in social networks using multiple-population mean field games. *IEEE Transactions on Signal and Information Processing over Networks*, 8:301–316, 2022.

Bowei Chen, Sen Xu, Heyang Xu, Xuesheng Bian, Naixuan Guo, Xiufang Xu, Xiaopeng Hua, and Tian Zhou. Structural deep multi-view clustering with integrated abstraction and detail. *Neural Networks*, 175:106287, 2024.

Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. Deep multiview clustering by contrasting cluster assignments. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16752–16761, 2023a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Zhe Chen, Xiao-Jun Wu, Tianyang Xu, and Josef Kittler. Fast self-guided multi-view subspace clustering. *IEEE transactions on image processing*, 32:6514–6525, 2023b.

Zhe Chen, Cheng Ma, Jun Huang, Tianyang Xu, and Xiao-Jun Wu. Bcn: Bidirectional contrastive learning net for multi-view clustering. *IEEE Signal Processing Letters*, 2025a.

Zhe Chen, Xiao-Jun Wu, Tianyang Xu, Hui Li, and Josef Kittler. Deep discriminative multi-view clustering. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025b.

Zhe Chen, Xiao-Jun Wu, Tianyang Xu, Hui Li, and Josef Kittler. Multi-layer multi-level comprehensive learning for deep multi-view clustering. *Information Fusion*, 116:102785, 2025c.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pp. 1–9, 2009.

Jinrong Cui, Yuting Li, Han Huang, and Jie Wen. Dual contrast-driven deep multi-view clustering. *IEEE Transactions on Image Processing*, 2024.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19440–19451, 2023.

Zhibin Dong, Meng Liu, Siwei Wang, Ke Liang, Yi Zhang, Suyuan Liu, Jiaqi Jin, Xinwang Liu, and En Zhu. Enhanced then progressive fusion with view graph for multi-view clustering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15518–15527, 2025.

Ran Eisenberg, Jonathan Svirsky, and Ofir Lindenbaum. COPER: Correlation-based permutations for multi-view clustering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=5ZEbpBYGwH.

Si-Guo Fang, Dong Huang, Xiao-Sha Cai, Chang-Dong Wang, Chaobo He, and Yong Tang. Efficient multi-view clustering via unified and discrete bipartite graph learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):11436–11447, 2023a.

Uno Fang, Man Li, Jianxin Li, Longxiang Gao, Tao Jia, and Yanchun Zhang. A comprehensive survey on multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(12): 12350–12368, 2023b.

Lunke Fei, Junlin He, Qi Zhu, Shuping Zhao, Jie Wen, and Yong Xu. Deep multi-view contrastive clustering via graph structure awareness. *IEEE Transactions on Image Processing*, 2025.

Junlong Ke, Zichen Wen, Yechenhao Yang, Chenhang Cui, Yazhou Ren, Xiaorong Pu, and Lifang He. Integrating vision-language semantic graphs in multi-view clustering. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 4273–4281, 2024.

Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 8547–8555, 2021.

Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11174–11183, 2021.

Omid Madani, Manfred Georg, and David A Ross. On using nearly-independent feature families for high precision and confidence. In *Asian conference on machine learning*, pp. 269–284. PMLR, 2012.

Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Zhenqiu Shu, Teng Sun, Yunwei Luo, and Zhengtao Yu. Ambiguous instance-aware contrastive network with multi-level matching for multi-view document clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20479–20487, 2025.

Chang Tang, Xinwang Liu, Xinzhong Zhu, En Zhu, Zhigang Luo, Lizhe Wang, and Wen Gao. Cgd: Multi-view clustering via cross-view graph diffusion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5924–5931, 2020.

Daniel J Trosten, Sigurd Lokse, Robert Jenssen, and Michael Kampffmeyer. Reconsidering representation alignment for multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1255–1265, 2021.

Daniel J Trosten, Sigurd Løkse, Robert Jenssen, and Michael C Kampffmeyer. On the effects of self-supervision and contrastive alignment in deep multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23976–23985, 2023.

Xinhang Wan, Jiyuan Liu, Hao Yu, Qian Qu, Ao Li, Xinwang Liu, Ke Liang, Zhibin Dong, and En Zhu. Contrastive continual multiview clustering with filtered structural fusion. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International conference on machine learning*, pp. 1083–1092. PMLR, 2015.

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. *IEEE Transactions on Multimedia*, 26:9150–9162, 2024.

Baili Xiao, Zhibin Dong, Ke Liang, Suyuan Liu, Siwei Wang, Tianrui Liu, Xingchen Hu, En Zhu, and Xinwang Liu. Easemvc: Efficient dual selection mechanism for deep multi-view clustering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20716–20726, 2025.

Yuan Xie, Bingqian Lin, Yanyun Qu, Cuihua Li, Wensheng Zhang, Lizhuang Ma, Yonggang Wen, and Dacheng Tao. Joint deep multi-view learning for image clustering. *IEEE Transactions on Knowledge and Data Engineering*, 33(11):3594–3606, 2020.

Jie Xu, Yazhou Ren, Huayi Tang, Xiaorong Pu, Xiaofeng Zhu, Ming Zeng, and Lifang He. Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9234–9243, 2021.

Jie Xu, Yazhou Ren, Huayi Tang, Zhimeng Yang, Lili Pan, Yang Yang, Xiaorong Pu, Philip S Yu, and Lifang He. Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7470–7482, 2022a.

Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16051–16060, 2022b.

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. *Advances in neural information processing systems*, 36:1119–1131, 2023.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22957–22966, 2024.

Junpeng Xu, Min Meng, Jigang Liu, and Jigang Wu. Deep multi-view clustering with diverse and discriminative feature learning. *Pattern Recognition*, 161:111322, 2025.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19863–19872, 2023.

Weiqing Yan, Tingyu Yang, and Chang Tang. Self-supervised semantic soft label learning network for deep multi-view clustering. *IEEE Transactions on Multimedia*, 2025.

Yongbo Yu, Zhoumin Lu, Feiping Nie, Weizhong Yu, Zongcheng Miao, and Xuelong Li. Pseudo-label guided bidirectional discriminative deep multi-view subspace clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2025.