Quality-Aware Robust Multi-View Clustering for Heterogeneous Observation Noise

Peihan Wu
22321313@zju.edu.cn
Zhejiang University
Hangzhou, China

Guanjie Cheng*
chengguanjie@zju.edu.cn
Zhejiang University
Hangzhou, China

Yufei Tong
yufeitong@zju.edu.cn
Zhejiang University
Hangzhou, China

Meng Xi
ximeng@zju.edu.cn
Zhejiang University
Hangzhou, China

Shuiguang Deng
dengsg@zju.edu.cn
Zhejiang University
Hangzhou, China

Abstract

Deep multi-view clustering has achieved remarkable progress but remains vulnerable to complex noise in real-world applications. Existing noisy robust methods predominantly rely on a simplified binary assumption, treating data as either perfectly clean or completely corrupted. This overlooks the prevalent existence of heterogeneous observation noise, where contamination intensity varies continuously across data. To bridge this gap, we propose a novel framework termed Quality-Aware Robust Multi-View Clustering (QARMVC). Specifically, QARMVC employs an information bottleneck mechanism to extract intrinsic semantics for view reconstruction. Leveraging the insight that noise disrupts semantic integrity and impedes reconstruction, we utilize the resulting reconstruction discrepancy to precisely quantify fine-grained contamination intensity and derive instance-level quality scores. These scores are integrated into a hierarchical learning strategy: at the feature level, a quality-weighted contrastive objective is designed to adaptively suppress the propagation of noise; at the fusion level, a high-quality global consensus is constructed via quality-weighted aggregation, which is subsequently utilized to align and rectify local views via mutual information maximization. Extensive experiments on five benchmark datasets demonstrate that QARMVC consistently outperforms state-of-the-art baselines, particularly in scenarios with heterogeneous noise intensities.

CCS Concepts

• **Computing methodologies** → **Unsupervised learning**; *Neural networks*; • **Information systems** → *Multimedia and multimodal retrieval.*

Keywords

multi-view clustering, information bottleneck, contrastive learning, mutual information

1 Introduction

In recent years, with the rapid advancement of sensing, communication, and data acquisition technologies, multi-view data has become increasingly prevalent in a wide range of real-world applications, such as industrial anomaly detection, social recommendation [4], and urban health profiling and prediction [22]. Unlike single-view data, multi-view data describes the same object from multiple heterogeneous perspectives, where different views may capture distinct yet complementary characteristics. Such a multi-perspective description provides richer semantic cues and stronger structural information, making it possible to uncover latent patterns that are difficult to identify from any individual view alone. As a result, effectively exploiting the consistency and complementarity across multiple views has become a central issue in unsupervised representation learning and data mining.

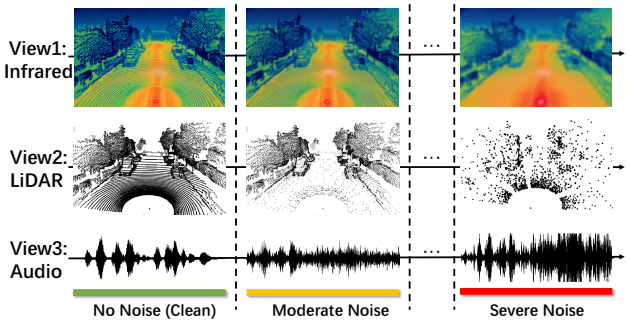


Figure 1: An illustrative diagram of heterogeneous noise intensity in a multi-view scenario. The diagram displays Infrared, LiDAR, and Audio views under varying environmental conditions. From left to right, the data quality exhibits a continuous degradation process from clean to severe noise, rather than a simple binary state.

To address this problem, Multi-View Clustering (MVC) has been widely studied as a fundamental unsupervised task, aiming to partition unlabeled data by jointly leveraging information from multiple views. Early MVC approaches were mainly built upon traditional machine learning paradigms, including multi-view subspace learning [9, 19], non-negative matrix factorization based methods [15, 25], and graph-based clustering frameworks [26, 31, 37]. These methods promote cross-view information fusion by constructing shared latent spaces, enforcing consistency constraints, or learning consensus graph structures. Although they have achieved encouraging results, their representation capacity is often limited by shallow architectures, hand-crafted similarity modeling, or relatively restrictive assumptions on data distributions. Benefiting

*Corresponding author.

from the strong nonlinear modeling ability of deep neural networks, Deep Multi-View Clustering (DMVC) has gradually become the dominant paradigm in this field [33] [41] [45] [3] [5]. Representative methods incorporate adversarial learning [23], contrastive feature learning [43] [36] [35] [2], and generative modeling [39] into multi-view clustering, substantially improving the quality of learned representations and the discriminability of clustering structures.

Despite the significant progress of existing DMVC methods, their robustness is still far from satisfactory when confronted with noisy real-world environments. Most existing studies model observation noise under a simplified binary assumption [42, 46], where each sample or view is regarded as either entirely clean or completely corrupted. However, such a coarse-grained assumption rarely holds in practice. Real-world multi-view data is more likely to suffer from heterogeneous observation noise, in which different instance-view pairs exhibit different and continuously varying levels of contamination [8]. As shown in Figure 1, sensory quality does not evolve in a discrete clean-versus-noisy manner, but instead spans a continuous spectrum from high-fidelity observations to mild degradation and even severe corruption. For example, in autonomous driving systems involving Infrared, LiDAR, and audio modalities, environmental disturbances such as adverse weather, motion blur, occlusion, and signal interference may affect different views to different extents. Under such conditions, some observations may still retain substantial semantic information despite moderate corruption, whereas others may become highly unreliable. This heterogeneous contamination pattern makes robust cross-view fusion substantially more difficult. Unfortunately, existing methods usually cannot explicitly characterize the fine-grained contamination intensity of each observation. Simply discarding non-ideal views as outliers may lead to the loss of useful complementary semantics, while blindly aggregating all views may introduce unreliable information into the common semantic space and deteriorate clustering quality. Therefore, how to accurately estimate the contamination intensity of each instance-view pair and conduct effective semantic learning under varying noise levels remains an urgent and largely underexplored problem in multi-view clustering.

To bridge this gap, we propose a **Q**uality-**A**ware **R**obust **M**ulti-**V**iew **C**lustering framework, termed **QARMVC**. Specifically, we employ the information bottleneck mechanism to capture intrinsic semantics by compressing each view into a compact latent space. Based on the insight that noise disrupts semantic integrity and impedes recovery, the resulting reconstruction discrepancy is utilized to precisely quantify heterogeneous contamination intensity, yielding fine-grained, instance-level quality scores. On this basis, we implement a hierarchical learning strategy. At the feature level, we utilize autoencoder-based embedding combined with a quality-weighted contrastive objective to ensure the semantic stability of the latent space. At the fusion level, view-specific embeddings are aggregated via quality-weighted fusion to construct a robust global consensus. We then maximize mutual information between this global target and local representations, effectively guiding noisy views to recover consistent semantics.

The key contributions of our paper are summarized as follows:

- We propose a novel Quality-Aware Robust Multi-View Clustering framework (QARMVC) to function robustly in heterogeneous noisy environments. To the best of our knowledge, this work is the first to incorporate fine-grained noise scores into multi-view clustering in a systematic manner.
- To perceive the heterogeneous contamination intensities, we introduce an information bottleneck mechanism to precisely quantify the quality of data. We further design a quality-weighted contrastive loss to ensure feature stability and construct a high-quality global representation to guide the rectification of contaminated views via mutual information maximization, thereby effectively alleviating noise interference.
- Extensive experiments on five benchmark datasets demonstrate that QARMVC consistently outperforms state-of-the-art baselines in both clustering accuracy and robustness under varying noise intensities.

2 Related Work

2.1 Deep Multi-View Clustering

Deep multi-view clustering (DMVC) has attracted increasing attention due to the strong representation learning capability of deep neural networks. Compared with traditional multi-view clustering methods based on subspace learning [9, 19], matrix factorization [15, 25], and graph fusion [31, 37], DMVC is able to learn nonlinear view-specific representations and cross-view interactions in an end-to-end manner, thereby showing stronger adaptability to complex data distributions. Existing DMVC methods can be roughly grouped into several representative lines. The first line follows clustering-oriented deep embedding, where feature learning and cluster structure optimization are jointly conducted in a unified framework. For example, DAMC [23] introduces adversarial learning to enhance the consistency of latent representations across views. The second line emphasizes contrastive representation learning, which improves clustering quality by strengthening agreement between semantically related samples or views. Representative methods such as [27, 43] exploit multi-level or decoupled contrastive objectives to enhance the discriminability and robustness of learned features. The third line focuses on graph-structured multi-view modeling, where relational information is explicitly encoded to preserve local topology and facilitate cross-view fusion [31, 37]. In addition, recent studies have further explored generative modeling for multi-view representation learning, such as diffusion-based frameworks that improve semantic completion and feature consistency under complex scenarios [39].

2.2 Noisy Multi-View Clustering

Despite the success of DMVC in handling multi-modal data, its performance often degrades in real-world applications due to complex noise. Generally, existing robust MVC research categorizes these challenges into three distinct scenarios: correspondence noise, missing noise, and absolute observation noise. Specifically, regarding correspondence noise, characterized by mismatches between samples, representative works [30] [29] [16] [12] tackle this misalignment by designing noise-tolerant contrastive losses or leveraging inter-view similarity contexts to rectify erroneous alignments. Missing noise deals with partial data loss, where imputation-based

methods [32] [44] [18] utilize observable neighboring samples to recover information, while non-imputation methods [40] [27] [48] rely on cross-view prediction in semantic space. Finally, absolute observation noise typically operates under a binary assumption, treating views as either entirely clean or completely corrupted [42]. Proactive methods address this by identifying and handling outliers. Specifically, AIRMVC [46] formulates this as an anomaly detection problem to distinguish clean samples from noisy outliers. RAC-DMVC [8] employs a cross-view cross-reconstruction mechanism, leveraging the information from reliable views to rectify the corrupted ones.

While the aforementioned studies have made significant strides, they overlook a more prevalent and realistic scenario: heterogeneous observation noise [38]. Existing methods for absolute noise have limited efficacy here, as they typically rely on at least one clean view per instance to identify and rectify noisy counterparts. However, in heterogeneous scenarios where data represents a mixture of valid semantics and noise, this binary treatment is ineffective: simply regarding such data as outliers overlooks the intrinsic semantic information, while indiscriminately fusing them leads to the distortion of the common semantic space. Therefore, bridging this gap requires a quality-aware framework capable of perceiving fine-grained noise intensities to balance semantic retention and noise suppression.

3 Methodology

3.1 Preliminary

Given a multi-view dataset $\mathcal{X} = \{X^1, X^2, \ldots, X^V\}$ consisting of N instances across V views, the data matrix for the v-th view is denoted as $X^v = [\mathbf{x}_1^v, \mathbf{x}_2^v, \ldots, \mathbf{x}_N^v] \in \mathbb{R}^{d_v \times N}$, where $\mathbf{x}_i^v \in \mathbb{R}^{d_v}$ represents the feature vector of the i-th instance in the v-th view. In real-world scenarios, the observed data typically suffers from heterogeneous observation noise, meaning that the dataset $\mathcal{X}$ consists of samples with non-uniform quality and varying levels of degradation. The primary objective is to partition the unlabeled dataset $\mathcal{X}$ into K disjoint clusters.

3.2 Quality Score Estimation

High-dimensional observations usually exhibit a low-dimensional intrinsic structure, enabling the extraction of core semantics by compressing the raw input into a compact latent variable. To derive a semantically rich representation, we construct an information bottleneck mechanism [1]. Our objective is to maximize the mutual information between the input view and its latent representation while constraining the capacity of the latent space:

$$\max I(X^v; \tilde{Z}) \quad \text{s.t.} \quad \dim(\tilde{Z}) \ll \dim(X^v). \tag{1}$$

Here, maximizing $I(X^v; \tilde{Z})$ ensures the preservation of intrinsic information from the original view X^v, while the dimensionality constraint prevents the latent variable from learning a trivial identity mapping. From an information-theoretic perspective, we first expand the mutual information term:

$$\begin{aligned} I(X^v; \tilde{Z}) &= \iint p(\mathbf{x}^v, \tilde{\mathbf{z}}) \log \frac{p(\mathbf{x}^v|\tilde{\mathbf{z}})}{p(\mathbf{x}^v)} \, d\mathbf{x}^v \, d\tilde{\mathbf{z}} \\ &= \iint p(\mathbf{x}^v) p(\tilde{\mathbf{z}}|\mathbf{x}^v) \log p(\mathbf{x}^v|\tilde{\mathbf{z}}) \, d\mathbf{x}^v \, d\tilde{\mathbf{z}} + H(X^v) \\ &\geq \iint p(\mathbf{x}^v) p(\tilde{\mathbf{z}}|\mathbf{x}^v) \log p(\mathbf{x}^v|\tilde{\mathbf{z}}) \, d\mathbf{x}^v \, d\tilde{\mathbf{z}}, \end{aligned} \tag{2}$$

where $\mathbf{x}^v$ and $\tilde{\mathbf{z}}$ are instances of X^v and $\tilde{Z}$, respectively. Since the true conditional distribution $p(\mathbf{x}^v|\tilde{\mathbf{z}})$ is unknown, we approximate it using a stochastic decoder $q^v(\mathbf{x}^v|\tilde{\mathbf{z}})$, which generates the reconstructed data $\tilde{\mathbf{x}}^v$. We further decompose the integral in Eq. (2) to derive a tractable variational lower bound:

$$\begin{aligned} I(X^v; \tilde{Z}) &\geq \iint p(\mathbf{x}^v) p(\tilde{\mathbf{z}}|\mathbf{x}^v) \log q^v(\mathbf{x}^v|\tilde{\mathbf{z}}) \, d\mathbf{x}^v \, d\tilde{\mathbf{z}} \\ &\quad + \iint p(\mathbf{x}^v) p(\tilde{\mathbf{z}}|\mathbf{x}^v) \log \frac{p(\mathbf{x}^v|\tilde{\mathbf{z}})}{q^v(\mathbf{x}^v|\tilde{\mathbf{z}})} \, d\mathbf{x}^v \, d\tilde{\mathbf{z}} \\ &\geq \mathbb{E}_{\mathbf{x}^v \sim p(\mathbf{x}^v)} \Big[\int p(\tilde{\mathbf{z}}|\mathbf{x}^v) \log q^v(\mathbf{x}^v|\tilde{\mathbf{z}}) \, d\tilde{\mathbf{z}} \Big]. \end{aligned} \tag{3}$$

By approximating the latent distribution $p(\tilde{\mathbf{z}}|\mathbf{x}^v)$ via a stochastic encoder, maximizing this lower bound is equivalent to minimizing the negative log-likelihood loss:

$$\mathcal{L}_{IB}^v = -\mathbb{E}_{\mathbf{x}^v \sim p(\mathbf{x}^v)} \Big[\mathbb{E}_{\tilde{\mathbf{z}} \sim p(\tilde{\mathbf{z}}|\mathbf{x}^v)} \big[\log q^v(\mathbf{x}^v|\tilde{\mathbf{z}}) \big] \Big]. \tag{4}$$

After training, clean samples adhering to the intrinsic manifold can be accurately reconstructed. Conversely, contaminated samples, disrupted by stochastic noise, fail to yield valid latent representations through the compression bottleneck, leading to significant reconstruction discrepancies. We quantify this contamination using the instance-level reconstruction error $R_i^v = \|\mathbf{x}_i^v - \tilde{\mathbf{x}}_i^v\|_1$. To facilitate view-adaptive weighting, we define a normalized contamination score:

$$C_i^v = \frac{R_i^v - \min(R^v)}{\max(R^v) - \min(R^v)}, \tag{5}$$

where $\max(R^v)$ and $\min(R^v)$ denote the extreme reconstruction errors computed across all samples in view v. The final quality score is obtained as $Q_i^v = (1 - C_i^v)^2$. This score acts as a dynamic weighting factor in the subsequent joint training phase, adaptively suppressing the influence of low-quality data.

3.3 Multi-view Representation Learning

With the estimated quality scores, we employ independent deep autoencoders to extract clustering-friendly features for each view. For the v-th view, the encoder $\mathcal{E}^v$ maps the input $\mathbf{x}_i^v$ into a latent representation $\mathbf{z}_i^v = \mathcal{E}^v(\mathbf{x}_i^v)$, while the decoder $\mathcal{D}^v$ generates the reconstruction $\hat{\mathbf{x}}_i^v = \mathcal{D}^v(\mathbf{z}_i^v)$. To ensure effective feature extraction, the model minimizes the aggregate reconstruction loss:

$$\mathcal{L}_{REC} = \sum_{v=1}^{V} \sum_{i=1}^{N} \|\mathbf{x}_i^v - \hat{\mathbf{x}}_i^v\|_2^2. \tag{6}$$

Building upon these latent representations, we introduce a contrastive learning mechanism to mitigate cross-view heterogeneity and capture consistent semantics. The semantic similarity between

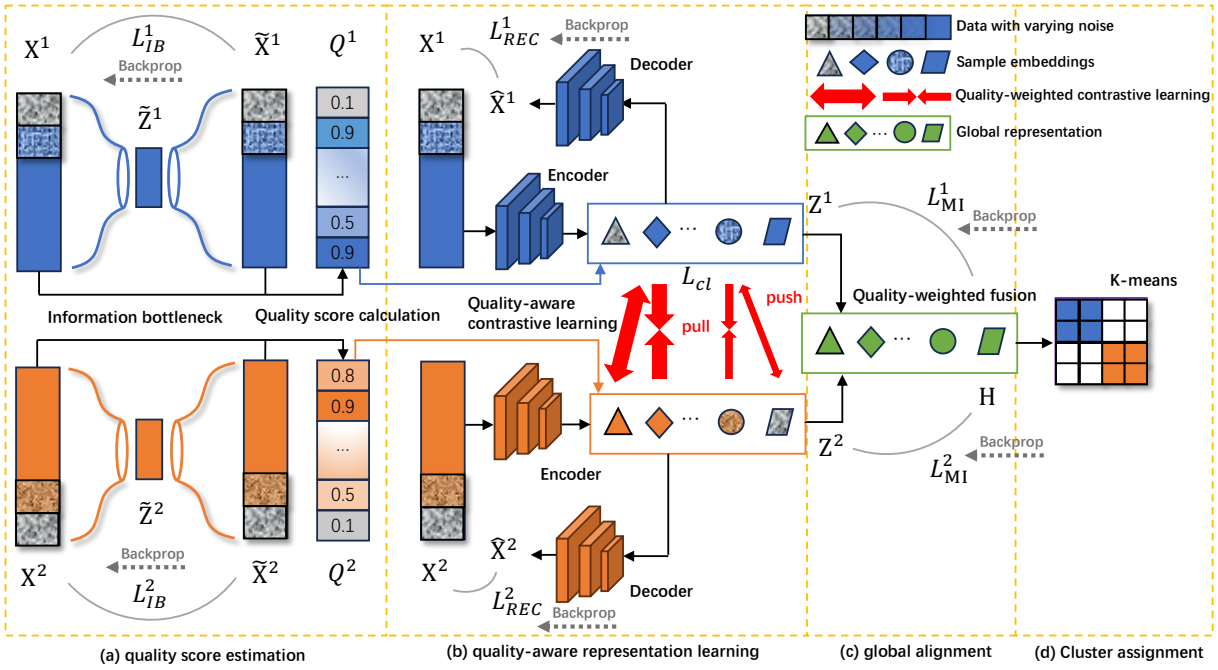


Figure 2: The framework consists of four modules: (a) Quality score estimation, where an information bottleneck mechanism quantifies heterogeneous noise intensity to derive instance-view specific quality scores; (b) Quality-aware representation learning, which utilizes these scores to re-weight contrastive learning, thereby suppressing noisy anchors; (c) Global alignment, where local views are aligned with a robust high-quality global consensus via mutual information maximization; (d) Cluster assignment: Perform K-means clustering on the global consensus representation H to obtain the final cluster labels.

representations is quantified using cosine similarity:

$$S(\mathbf{z}_i^u, \mathbf{z}_j^v) = \frac{(\mathbf{z}_i^u)^\top \mathbf{z}_j^v}{\|\mathbf{z}_i^u\|_2 \|\mathbf{z}_j^v\|_2}. \tag{7}$$

To align information across perspectives, we formulate an instance-level contrastive objective. Specifically, taking $\mathbf{z}_i^u$ from view u as the anchor and view v ($u \neq v$) as the target, the contrastive loss for the i-th instance is defined as:

$$\ell_i^{(u\to v)} = -\log \frac{\exp(S(\mathbf{z}_i^u, \mathbf{z}_i^v)/\tau)}{\sum_{k=1}^N \exp(S(\mathbf{z}_i^u, \mathbf{z}_k^v)/\tau)}, \tag{8}$$

where τ denotes the temperature hyperparameter. Minimizing Eq. (8) effectively pulls the positive counterpart $\mathbf{z}_i^v$ toward the anchor $\mathbf{z}_i^u$ in the latent space, while simultaneously pushing the negative samples $\{\mathbf{z}_k^v\}_{k\neq i}$ away.

Standard contrastive frameworks typically treat all anchors equally by minimizing the average of $\ell_i^{(u\to v)}$. However, such an indiscriminate strategy is suboptimal in the presence of heterogeneous observation noise; aligning corrupted anchors with other views inevitably propagates noise and distorts the common semantic space. To address this, we propose a quality-aware robust contrastive loss by incorporating the instance-view quality score Q_i^u. By adaptively re-weighting the contribution of each anchor based on its reliability, the final objective is formulated as:

$$\mathcal{L}_{RCL} = \sum_{i=1}^N \sum_{u=1}^V \sum_{v\neq u} Q_i^u \cdot \ell_i^{(u\to v)}. \tag{9}$$

This formulation ensures that high-quality instances dominate the semantic alignment process, while the negative impact of contaminated data is suppressed, facilitating the learning of robust, view-consistent representations.

3.4 Quality-Guided Global Fusion and Alignment

To leverage cross-view complementarity while maintaining robustness, we introduce a global-local alignment module consisting of quality-guided fusion and mutual information (MI) maximization [24]. Standard fusion strategies [7], such as simple concatenation or averaging, treat all views indiscriminately, which often leads to the contamination of the common semantic space by noisy observations. To address this, we utilize the estimated quality scores to construct a robust global representation.

Specifically, we first normalize the instance-view quality scores to obtain the fusion weights $w_i^v = Q_i^v / \sum_{k=1}^V Q_i^k$. Subsequently, the global consensus representation $\mathbf{h}_i$ for the i-th instance is generated via a weighted aggregation of view-specific embeddings:

$$\mathbf{h}_i = \sum_{v=1}^V w_i^v \mathbf{z}_i^v, \tag{10}$$

where $\mathbf{z}_i^v$ denotes the embedding of the i-th instance in the v-th view. By prioritizing high-quality views via Eq. (10), the global representation H successfully extracts representative semantics while adaptively suppressing noise interference.

Algorithm 1 Training Procedure of QARMVC

Require: Multi-view dataset $\mathcal{X} = \{X^v\}_{v=1}^V$, cluster number K, max epochs E
Ensure: Clustering results $\mathbf{y}$

```
Train the information bottleneck module by minimizing Eq. (4);
Compute the contamination scores via Eq. (5) and obtain the quality scores Q_i^v = (1 − C_i^v)^2;
for epoch = 1 to E do
    Encode each view to obtain latent representations {Z^v}_{v=1}^V;
    Compute the reconstruction loss L_REC via Eq. (6);
    Compute the quality-aware contrastive loss L_RCL via Eq. (9);
    Construct the global consensus representation H via Eq. (10);
    Compute the mutual-information alignment loss L_MI via Eq. (11);
    Update network parameters by minimizing Eq. (12);
end for
Obtain the final cluster labels y by performing KMeans on the global consensus representation H.
return y
```

To further utilize this robust global consensus to guide and rectify view-specific learning, we maximize the consistency between H and each local representation Z^v. This objective is formulated as minimizing the negative mutual information loss:

$$\begin{aligned}\mathcal{L}_{MI} &= -\sum_{v=1}^{V} I(H; Z^v) \\ &= -\sum_{v=1}^{V}\sum_{i,j} p(\mathbf{h}_i, \mathbf{z}_j^v) \log \frac{p(\mathbf{h}_i, \mathbf{z}_j^v)}{p(\mathbf{h}_i)p(\mathbf{z}_j^v)},\end{aligned} \tag{11}$$

where $p(\mathbf{h}_i, \mathbf{z}_j^v)$ denotes the joint probability distribution, and $p(\mathbf{h}_i)$, $p(\mathbf{z}_j^v)$ are the corresponding marginals.

Minimizing $\mathcal{L}_{MI}$ aligns view-specific features with the global consensus. This high-quality target provides reliable semantic guidance to rectify local inconsistencies and distortions in contaminated views, thereby fostering a unified and robust latent space on which the final clustering is performed.

3.5 Overall Algorithm

The comprehensive objective function of QARMVC integrates quality-weighted reconstruction, robust contrastive learning and mutual information alignment. The total loss is formulated as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{REC} + \lambda_2 \mathcal{L}_{RCL} + \lambda_3 \mathcal{L}_{MI}, \tag{12}$$

where λ_1, λ_2 and λ_3 are trade-off hyperparameters. The optimization process, detailed in Algorithm 1.

4 Experiments

To comprehensively evaluate QARMVC, we conduct experiments from six aspects. First, we compare QARMVC with several representative deep multi-view clustering methods to assess its overall clustering performance. Second, we analyze the effectiveness of the proposed information bottleneck based quality estimation module by examining the correlation between the estimated scores and the actual contamination intensity. Third, we perform ablation studies to verify the contribution of each core component. Fourth, we visualize the learned latent representations and clustering structures to further illustrate the representation quality of QARMVC. Fifth, we investigate the sensitivity of QARMVC to key hyper-parameters, including the loss trade-off coefficients and the bottleneck dimension. Finally, we evaluate QARMVC in a comprehensive real-world scenario to examine its practical effectiveness under naturally existing quality variations.

4.1 Experimental Settings

Datasets: We conduct experiments on five widely used multi-view benchmark datasets. Scene-15 [6] comprises 4,485 images belonging to 15 classes, encompassing both indoor and outdoor environments. For each image, we extract GIST and PHOG features. MNIST-USPS [21] consists of 5,000 handwritten digit images distributed across 10 digit categories. We utilize feature vectors from MNIST and USPS sources. LandUse-21 [47] contains 2,100 satellite imagery samples categorized into 21 classes. We employ PHOG and LBP features to represent the visual information. ALOI [10] is a collection of 10,800 object images belonging to 100 categories. We adopt Color Similarity and HSV histograms. Finally, 100leaves [49] contains 1,600 samples from 100 categories. We extract FSM and SD features for analysis.

To simulate heterogeneous observation noise in real-world scenarios, we follow the noise injection protocol commonly used in existing robust multi-view clustering methods [8, 42], and further extend it to model fine-grained variations in noise severity. Specifically, a proportion of samples is randomly selected for contamination according to $\eta \in \{10\%, 30\%, 50\%\}$. For each selected sample, Gaussian noise is injected with an intensity coefficient $\alpha \in \{0.2, 0.4, \ldots, 1.0\}$. Accordingly, the contaminated view $\bar{x}$ is constructed as $\bar{x} = \alpha \cdot \delta + (1 - \alpha) \cdot x$, where δ denotes Gaussian noise.

Comparison Algorithms: To showcase the broad applicability and superior performance of QARMVC, we compare it against several state-of-the-art deep multi-view clustering methods, including noise-resilient methods designed to handle corrupted data (CANDY [11], DIVIDE [27], RAC-DMVC [8], and MVCAN [42]) and classical methods (MSDIB [13], STCMC_UR [14] and SURE [44]). For fair comparison, all baselines are reproduced using their official implementations.

Implementation Details. In this study, all experiments are implemented using the PyTorch framework [17] on a workstation equipped with a single NVIDIA RTX 4090 GPU. The proposed QARMVC is trained in an end-to-end manner using the Adam optimizer [20], with the learning rate fixed at 1×10^{-3}. The total training process spans 200 epochs with a batch size of 128. As for the hyperparameters in the overall objective function Eq. (12), the trade-off weights are consistently set as $\lambda_1 = 0.1$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.1$.

4.2 Comparison Experiments

In this subsection, we compare QARMVC with several representative baselines under different heterogeneous observation noise

Table 1: Clustering performance comparison on five benchmark datasets under varying heterogeneous observation noise ratios.

Ratio	Methods	Scene15			MNIST-USPS			LandUse21			ALOI			100Leaves		
		ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
10%	SURE	38.57	38.99	23.00	90.98	87.05	82.21	24.19	25.86	9.54	**59.26**	78.70	43.28	42.75	76.07	24.20
	CANDY	38.10	36.96	21.87	75.86	74.76	66.15	25.62	30.03	12.28	17.24	56.72	17.59	51.69	75.11	36.86
	DIVIDE	38.95	36.30	20.53	78.86	80.11	70.72	25.14	29.14	11.34	19.32	63.28	21.98	70.86	86.42	64.68
	MVCAN	26.91	27.24	14.09	56.30	51.87	38.75	20.33	23.77	7.46	40.83	64.77	31.09	72.81	85.51	60.46
	RAC-DMVC	40.32	40.11	**24.60**	78.38	68.30	73.17	24.90	29.44	**12.31**	20.54	64.24	22.19	64.88	79.51	48.57
	MSDIB	34.22	34.48	19.32	61.08	65.72	51.13	23.85	28.84	10.66	45.46	69.99	33.18	65.63	80.84	50.52
	STCMC_UR	30.16	37.72	20.58	84.72	85.87	77.88	18.85	26.64	6.67	27.18	50.44	4.28	66.78	79.39	49.84
	QARMVC	**43.83**	**40.32**	24.21	**96.54**	**90.14**	**91.45**	**25.72**	**31.66**	12.03	49.05	**78.75**	**49.65**	**77.44**	**89.33**	**69.45**
30%	SURE	35.67	38.58	21.19	80.64	71.45	65.20	22.03	21.34	8.03	41.56	70.38	34.46	44.19	71.63	27.44
	CANDY	32.40	28.20	15.96	66.50	72.24	61.86	24.43	23.66	8.93	17.63	57.24	18.20	38.25	65.95	24.38
	DIVIDE	38.19	35.51	19.54	78.76	67.85	58.58	24.29	29.33	10.79	17.89	58.59	18.80	58.13	76.91	43.90
	MVCAN	26.73	24.30	13.02	49.86	46.87	32.64	15.14	16.19	4.21	12.75	30.59	6.45	54.31	74.74	38.28
	RAC-DMVC	30.26	28.08	13.94	70.62	53.29	59.86	22.38	22.68	8.30	17.23	58.12	18.70	50.06	70.78	33.21
	MSDIB	28.33	27.25	12.70	52.20	51.58	35.12	18.42	21.72	6.99	20.09	43.00	10.50	51.42	69.35	32.58
	STCMC_UR	27.67	34.33	15.09	79.26	83.38	74.60	17.95	25.51	6.03	18.50	42.78	3.02	50.15	69.75	32.73
	QARMVC	**42.72**	**40.14**	**23.97**	**94.73**	**88.32**	**88.83**	**24.82**	**29.52**	**11.03**	**42.60**	**71.86**	**41.15**	**70.94**	**84.89**	**60.05**
50%	SURE	30.42	30.71	15.81	64.92	58.82	49.80	15.57	14.15	3.73	20.50	44.74	7.99	29.88	60.47	13.67
	CANDY	25.33	20.92	10.38	68.04	54.70	48.00	15.62	13.26	3.53	10.42	39.65	7.02	27.87	57.69	13.13
	DIVIDE	29.14	27.10	13.17	73.28	60.61	48.59	19.10	21.02	6.50	11.06	41.86	7.93	43.31	67.46	26.56
	MVCAN	19.60	14.92	7.64	47.02	44.27	26.83	12.13	12.03	3.98	5.99	16.17	1.35	41.94	67.71	25.15
	RAC-DMVC	30.26	28.08	13.94	50.96	31.20	42.03	16.71	15.26	4.36	11.32	41.24	8.02	38.31	64.17	23.08
	MSDIB	22.67	23.91	9.91	33.56	32.47	17.14	16.76	17.12	4.79	5.83	20.03	1.72	39.42	63.55	23.88
	STCMC_UR	25.68	25.19	9.53	56.42	64.38	44.58	14.80	18.65	3.16	12.37	31.10	1.13	40.34	63.23	23.18
	QARMVC	**35.42**	**33.63**	**19.12**	**93.74**	**85.69**	**86.33**	**22.45**	**24.84**	**9.02**	**23.12**	**50.02**	**16.75**	**54.19**	**75.35**	**38.62**

Bold indicates the best performance, and underline indicates the second best.

ratios. The quantitative results are reported in Table 1, where the best and second-best results are highlighted in bold and underlined, respectively. Overall, QARMVC achieves the best or highly competitive performance on most dataset-metric pairs, demonstrating its effectiveness in noisy multi-view clustering.

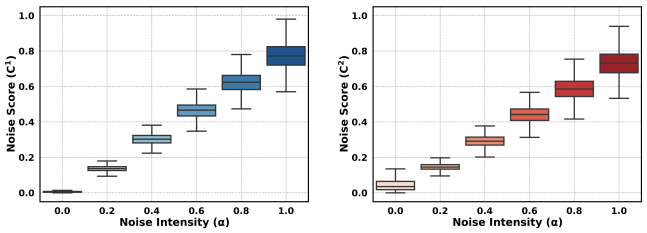
As the noise ratio increases from 10% to 50%, the performance of most baselines degrades substantially, whereas QARMVC remains much more stable. This result suggests that the proposed quality-aware framework can effectively suppress the influence of unreliable observations and preserve more discriminative clustering structures under severe heterogeneous noise.

In particular, on MNIST-USPS, QARMVC consistently achieves the best results across all three noise settings, reaching 96.54/90.14/91.45, 94.73/88.32/88.83, and 93.74/85.69/86.33 in terms of ACC/NMI/ARI, respectively. Under 50% noise, it improves ACC over the strongest baseline by 20.46 percentage points, which clearly verifies its robustness in heavily corrupted scenarios. Similar improvements can also be observed on Scene15 and ALOI, especially under medium and high noise ratios.

Table 2: Analysis between noise scores(C^v) and intensities(α).

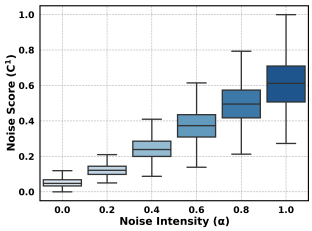
Dataset	Feature	BottleneckDim	Pearson	Spearman
Scene15	GIST	7	0.933	0.904
	PHOG	20	0.979	0.925
MNIST-USPS	MNIST	520	0.921	0.779
	USPS	100	0.927	0.824
LandUse-21	PHOG	7	0.844	0.834
	LBP	20	0.977	0.920
ALOI	CS	20	0.991	0.932
	HSV	20	0.979	0.921
100Leaves	FSM	32	0.943	0.882
	SD	32	0.917	0.822

Although QARMVC is not always ranked first on a few relatively easy cases, it shows a clear advantage as the noise level becomes higher. In summary, these results demonstrate that QARMVC not only yields strong clustering performance, but also exhibits superior robustness against heterogeneous observation noise.

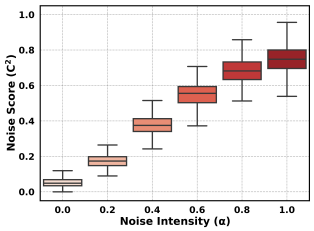


(a) ALOI: Color Similarity

(b) ALOI: HSV Histogram



(c) Scene15: GIST



(d) Scene15: PHOG

Figure 3: Noise score analysis on the ALOI and Scene15 datasets.

4.3 Quality Score Analysis

To validate the capability of QARMVC in accurately perceiving contamination intensity, we conduct a quantitative analysis under the 50% noise ratio setting. Specifically, Figure 3 presents the distributions of the estimated noise scores (C^v) with respect to different noise intensities (α) on the ALOI and Scene15 datasets. As the noise intensity increases, the estimated scores in all views exhibit a clear upward trend, and the corresponding box plots show a strong monotonic relationship between the estimated contamination level and the actual corruption intensity. Moreover, no disproportionately extreme outliers are observed in the estimated score distributions, suggesting that the proposed estimator retains sufficient resolution across different noise levels. This observation is further supported by the quantitative results in Table 2, where consistently high Pearson and Spearman correlation coefficients are achieved across different datasets and feature views. These results confirm that the proposed estimator can effectively capture contamination intensity, making $Q^v = (1 - C^v)^2$ a reliable measure of data quality for quality-aware learning and fusion.

4.4 Ablation Study

Table 3 reports the ablation results on MNIST-USPS and 100leaves under 50% noise. The full QARMVC achieves the best performance on both datasets, demonstrating the effectiveness of jointly optimizing reconstruction, quality-aware contrastive learning, and cross-view alignment. Among these components, the quality-aware contrastive objective is the most critical, as removing $\mathcal{L}_{RCL}$ causes a dramatic performance drop on both datasets. Removing $\mathcal{L}_{REC}$ or $\mathcal{L}_{MI}$ also consistently degrades the results, indicating that preserving view-specific information and enforcing semantic consistency are both necessary for robust fusion. In addition, QARMVC consistently outperforms Standard CL, which verifies the advantage of the proposed quality-aware contrastive strategy over indiscriminate pairwise alignment. The much weaker performance of the double-ablation variants further suggests that no single objective is sufficient, and the superiority of the full model comes from the cooperation of all components.

Table 3: Ablation study under 50% noise ratio.

Settings	MNIST-USPS			100leaves		
	ACC	NMI	ARI	ACC	NMI	ARI
w/o $\mathcal{L}_{RCL}$ & $\mathcal{L}_{MI}$	28.32	25.45	12.88	40.34	73.55	24.01
w/o $\mathcal{L}_{REC}$ & $\mathcal{L}_{MI}$	78.23	70.44	68.55	52.50	74.24	37.23
w/o $\mathcal{L}_{REC}$ & $\mathcal{L}_{RCL}$	28.78	26.56	12.22	40.13	73.13	23.54
w/o $\mathcal{L}_{MI}$	88.85	82.64	79.77	51.33	73.53	35.23
w/o $\mathcal{L}_{RCL}$	30.12	27.33	13.58	41.81	68.47	24.89
w/o $\mathcal{L}_{REC}$	80.54	73.58	69.11	52.38	74.65	37.78
Standard CL	88.23	81.34	78.22	50.19	73.30	34.22
QARMVC (Full)	**93.74**	**85.69**	**86.33**	**54.19**	**75.35**	**38.62**

Bold indicates the best performance,

4.5 Parameter Analysis

Loss function hyperparameter analysis. As shown in Figure 4, QARMVC is relatively robust to the choices of λ_1, λ_2, and λ_3 on MNIST-USPS with 10% noise. With respect to the interaction between λ_1 and λ_2, better results are generally obtained when λ_2 is set to a moderate value, while overly small λ_2 leads to clear performance degradation, indicating that the quality-aware contrastive objective should maintain sufficient influence during training. A similar trend can also be observed for the combination of λ_1 and λ_3, where the model achieves stronger performance in the middle parameter region, whereas extreme settings tend to weaken clustering quality. Overall, the best results are obtained by balanced parameter combinations rather than boundary values, which suggests that QARMVC does not rely on delicate hyperparameter tuning and maintains stable performance over a reasonably wide range.

Bottleneck dimension analysis. We further study the effect of the bottleneck dimension on the quality estimation module in Figure 6. On both ALOI and Scene15, the Pearson and Spearman correlation coefficients remain consistently high across a broad range of bottleneck ratios, indicating that the proposed estimator is not overly sensitive to this hyperparameter and thus enjoys satisfactory robustness. Meanwhile, the best or near-best performance is generally achieved in the middle region. In contrast, overly small bottleneck dimensions may lead to insufficient information preservation, while excessively large ones weaken the desired compactness of the latent representation. Therefore, we recommend setting the bottleneck dimension to approximately 1/2 of the input feature dimension in practice.

4.6 Visualization Analysis

To intuitively evaluate the learned representations, we visualize the latent embedding spaces of QARMVC and representative baselines on the MNIST-USPS dataset (10% noise) via t-SNE [34] in Figure 5. While baseline methods exhibit blurred boundaries and considerable overlap indicating vulnerability to noise, QARMVC

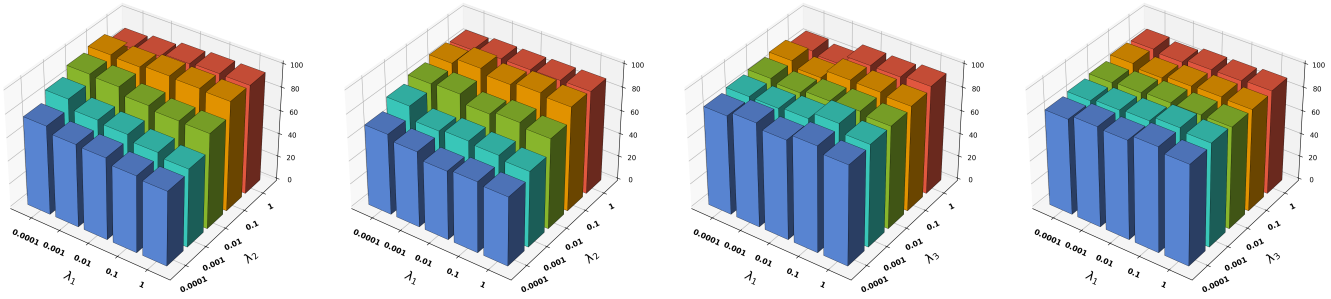


Figure 4: Sensitivity analysis on MNIST-USPS dataset with 10% Noise.

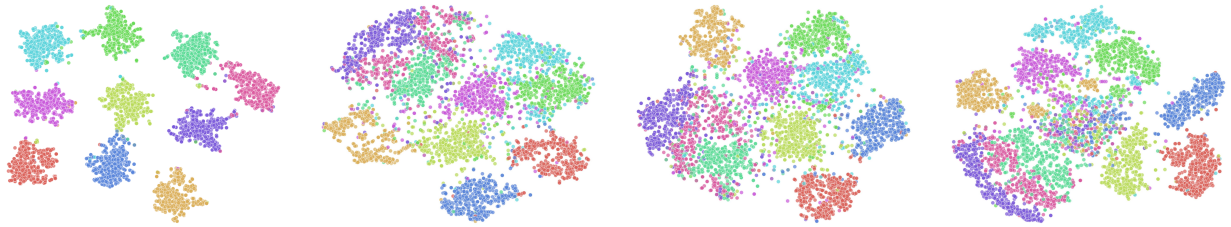


Figure 5: Visualization on MNIST-USPS dataset with 10% Noise.

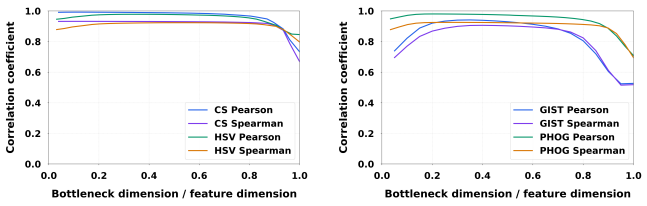


Figure 6: Correlation analysis of estimated noise scores under different bottleneck dimensions.

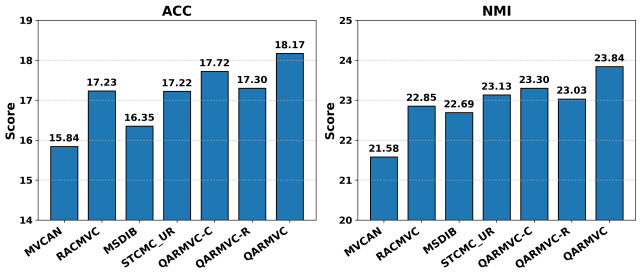


Figure 7: ACC and NMI comparison on the SUNRGBD dataset. QARMVC-C and QARMVC-R denote the variants of QARMVC using constant and random scores, respectively.

generates a highly discriminative latent space characterized by high intra-cluster compactness and clear inter-cluster separability. This distinct structure confirms that our quality-aware framework effectively filters out noise interference to yield robust and consistent semantics, significantly surpassing competing methods.

4.7 Application to Comprehensive Scenario

We further evaluate QARMVC on the SUNRGBD dataset [28] without following the noise injection protocol, in order to verify its effectiveness in a comprehensive real-world scenario.

SUNRGBD is a representative RGB-D scene understanding dataset collected from real indoor environments. With multi-modal observations and substantial real-world variations, such as illumination changes, viewpoint shifts, background clutter, object occlusion, and sensor-induced quality fluctuations, it provides a challenging benchmark for robust multi-view clustering. We adopt the two-view setting and directly evaluate all methods on the original data.

As shown in Fig. 7, QARMVC achieves the best ACC and NMI, reaching 18.17 and 24.49, respectively. It outperforms the strongest baseline by 0.94 points in ACC and 1.36 points in NMI, which suggests its effectiveness under naturally existing quality variations.

When the learned quality score is replaced by a constant or random one, the performance consistently drops to 17.72/23.30 and 17.30/23.03 in terms of ACC/NMI, respectively. This confirms that the improvement is brought by informative quality estimation rather than naive weighting. These results provide preliminary evidence of the practical applicability of QARMVC on real-world multi-view data.

5 Conclusion

In this paper, we identify and address the challenge of heterogeneous observation noise in multi-view clustering. To this end, we propose Quality-Aware Robust Multi-View Clustering (QARMVC), a novel framework that leverages an information bottleneck mechanism to quantify instance-level data quality. These quality scores guide a weighted contrastive objective and a global-local alignment module, enabling the model to suppress noisy anchors and align distorted views through a robust global consensus. Extensive experiments on multiple benchmarks demonstrate the effectiveness of QARMVC, especially under varying noise intensities. Nevertheless, the current framework is mainly designed for unstructured random noise and may be less effective for more structured noise patterns. Exploring more adaptive robust multi-view clustering under broader real-world noise settings will be an important direction for future work.

References

[1] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. Deep Variational Information Bottleneck. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HyxQzBceg

[2] Ronggang Cai, Hongmei Chen, Yong Mi, Chuan Luo, Shi-Jinn Horng, and Tianrui Li. 2024. Multi-view Clustering via Pseudo-label Guide Learning and Latent Graph Structure Recovery. *Pattern Recognition* 151 (2024), 110420. doi:10.1016/j.patcog.2024.110420

[3] Jie Chen, Hua Mao, Wai Lok Woo, and Xi Peng. 2023. Deep Multiview Clustering by Contrasting Cluster Assignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16706–16715.

[4] Rui Chen, Jialu Chen, and Xianghua Gan. 2024. Multi-view graph contrastive learning for social recommendation. *Scientific reports* 14, 1 (2024), 22643.

[5] Jinrong Cui, Yuting Li, Han Huang, and Jie Wen. 2024. Dual Contrast-Driven Deep Multi-View Clustering. *IEEE Transactions on Image Processing* 33 (2024), 4753–4764. doi:10.1109/TIP.2024.3444269

[6] Dengxin Dai and Luc Van Gool. 2013. Ensemble projection for semi-supervised image classification. In *Proceedings of the IEEE international conference on computer vision*. 2072–2079.

[7] Xiaojian Ding, Lin Zhao, Xian Li, and Xiaoying Zhu. 2025. Incomplete Multi-view Clustering via Hierarchical Semantic Alignment and Cooperative Completion. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=GplW3hkvnr

[8] Shihao Dong, Yue Liu, Xiaotong Zhou, Yuhui Zheng, Huiying Xu, and Xinzhong Zhu. 2025. RAC-DMVC: Reliability-Aware Contrastive Deep Multi-View Clustering under Multi-Source Noise. *arXiv preprint arXiv:2511.13561* (2025).

[9] Hongchang Gao, Feiping Nie, Xuelong Li, and Heng Huang. 2015. Multi-view subspace clustering. In *Proceedings of the IEEE international conference on computer vision*. 4238–4246.

[10] Jan-Mark Geusebroek, Gertjan J Burghouts, and Arnold WM Smeulders. 2005. The Amsterdam library of object images. *International Journal of Computer Vision* 61, 1 (2005), 103–112.

[11] Ruiming Guo, Mouxing Yang, Yijie Lin, Xi Peng, and Peng Hu. 2024. Robust contrastive multi-view clustering against dual noisy correspondence. *Advances in Neural Information Processing Systems* 37 (2024), 121401–121421.

[12] Changhao He, Hongyuan Zhu, Peng Hu, and Xi Peng. 2024. Robust Variational Contrastive Learning for Partially View-unaligned Clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 4167–4176. doi:10.1145/3664647.3681331

[13] Shizhe Hu, Jiahao Fan, Guoliang Zou, and Yangdong Ye. 2025. Multi-aspect Self-guided Deep Information Bottleneck for Multi-modal Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 17314–17322.

[14] Shizhe Hu, Binyan Tian, Weibo Liu, and Yangdong Ye. 2025. Self-supervised Trusted Contrastive Multi-view Clustering with Uncertainty Refined. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 17305–17313.

[15] Shudong Huang, Zhao Kang, and Zenglin Xu. 2020. Auto-weighted multi-view clustering via deep matrix decomposition. *Pattern Recognition* 97 (2020), 107015.

[16] Zhenyu Huang, Peng Hu, Joey Tianyi Zhou, Jiancheng Lv, and Xi Peng. 2020. Partially View-aligned Clustering. In *Advances in Neural Information Processing Systems*, Vol. 33.

[17] Sagar Imambi, Kolla Bhanu Prakash, and GR Kanagachidambaresan. 2021. PyTorch. *Programming with TensorFlow: solution for edge computing applications* (2021), 87–104.

[18] Jiaqi Jin, Siwei Wang, Zhibin Dong, Xinwang Liu, and En Zhu. 2023. Deep incomplete multi-view clustering with cross-view partial sample and prototype alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11600–11609.

[19] Zhao Kang, Wangtao Zhou, Zhitong Zhao, Junming Shao, Meng Han, and Zenglin Xu. 2020. Large-Scale Multi-View Subspace Clustering in Linear Time. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 4412–4419. doi:10.1609/AAAI.V34I04.5867

[20] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[21] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[22] Jinlin Li and Xiao Zhou. 2025. CureGraph: Contrastive multi-modal graph representation learning for urban living circle health profiling and prediction. *Artificial Intelligence* 340 (2025), 104278.

[23] Zhaoyang Li, Qianqian Wang, Zhiqiang Tao, Quanxue Gao, and Zhaohua Yang. 2019. Deep adversarial multi-view clustering network. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)* (Macao, China). AAAI Press, 2952–2958.

[24] Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. 2022. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4447–4461.

[25] Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han. 2013. Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of the 2013 SIAM international conference on data mining*. SIAM, 252–260.

[26] Suyuan Liu, Siwei Wang, Pei Zhang, Kai Xu, Xinwang Liu, Changwang Zhang, and Feng Gao. 2022. Efficient one-pass multi-view subspace clustering with consensus anchors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 7576–7584.

[27] Yiding Lu, Yijie Lin, Mouxing Yang, Dezhong Peng, Peng Hu, and Xi Peng. 2024. Decoupled contrastive multi-view clustering with high-order random walks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 14193–14201.

[28] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 567–576.

[29] Yuan Sun, Yongxiang Li, Zhenwen Ren, Guiduo Duan, Dezhong Peng, and Peng Hu. 2025. ROLL: Robust Noisy Pseudo-label Learning for Multi-View Clustering with Noisy Correspondence. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 30732–30741.

[30] Yuan Sun, Yang Qin, Yongxiang Li, Dezhong Peng, Xi Peng, and Peng Hu. 2024. Robust Multi-View Clustering with Noisy Correspondence. *IEEE Transactions on Knowledge and Data Engineering* (2024).

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

[32] Huayi Tang and Yong Liu. 2022. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *International conference on machine learning*. PMLR, 21090–21110.

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

[34] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[35] Fangdi Wang, Jiaqi Jin, Zhibin Dong, Xihong Yang, Yu Feng, Xinwang Liu, Xinzhong Zhu, Siwei Wang, Tianrui Liu, and En Zhu. 2024. View Gap Matters: Cross-view Topology and Information Decoupling for Multi-view Clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 8431–8440. doi:10.1145/3664647.3680915

[36] Fangdi Wang, Jiaqi Jin, Jingtao Hu, Suyuan Liu, Xihong Yang, Siwei Wang, Xinwang Liu, and En Zhu. 2024. Evaluate then Cooperate: Shapley-based View Cooperation Enhancement for Multi-view Clustering. In *Advances in Neural Information Processing Systems*, Vol. 37.

[37] Hao Wang, Yan Yang, and Bing Liu. 2019. GMC: Graph-based multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering* 32, 6 (2019), 1116–1129.

[38] Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, and Chengjie Wang. 2023. Multimodal industrial anomaly detection via hybrid fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8032–8041.

[39] Jie Wen, Shijie Deng, Waikeung Wong, Guoqing Chao, Chao Huang, Lunke Fei, and Yong Xu. 2024. Diffusion-based missing-view generation with the application on incomplete multi-view clustering. In *Forty-First International Conference on Machine Learning*.

[40] Jie Xu, Chao Li, Yazhou Ren, Liang Peng, Yujie Mo, Xiaoshuang Shi, and Xiaofeng Zhu. 2022. Deep incomplete multi-view clustering via mining cluster complementarity. In *Proceedings of the AAAI conference on artificial intelligence*,

Vol. 36. 8761–8769.

[41] Jie Xu, Yazhou Ren, Huayi Tang, Zhimeng Yang, Lili Pan, Yang Yang, Xiaorong Pu, Philip S. Yu, and Lifang He. 2023. Self-Supervised Discriminative Feature Learning for Deep Multi-View Clustering. *IEEE Transactions on Knowledge and Data Engineering* 35, 7 (2023), 7470–7482. doi:10.1109/TKDE.2022.3193569

[42] Jie Xu, Yazhou Ren, Xiaolong Wang, Lei Feng, Zheng Zhang, Gang Niu, and Xiaofeng Zhu. 2024. 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*. 22957–22966.

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

[44] Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. 2022. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2022), 1055–1069.

[45] Xihong Yang, Jiaqi Jin, Siwei Wang, Ke Liang, Yue Liu, Yi Wen, Suyuan Liu, Sihang Zhou, Xinwang Liu, and En Zhu. 2023. DealMVC: Dual Contrastive Calibration for Multi-view Clustering. In *Proceedings of the 31st ACM International Conference on Multimedia*. 337–346. doi:10.1145/3581783.3611951

[46] Xihong Yang, Siwei Wang, Fangdi Wang, Jiaqi Jin, Suyuan Liu, Yue Liu, En Zhu, Xinwang Liu, and Yueming Jin. 2025. Automatically Identify and Rectify: Robust Deep Contrastive Multi-view Clustering in Noisy Scenarios. In *Forty-second International Conference on Machine Learning*. https://openreview.net/forum?id=iFOXz5H2gB

[47] Yi Yang and Shawn Newsam. 2010. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*. 270–279.

[48] Yuanyang Zhang, Yijie Lin, Weiqing Yan, Li Yao, Xinhang Wan, Guangyuan Li, Chao Zhang, Guanzhou Ke, and Jie Xu. 2025. Incomplete Multi-view Clustering via Diffusion Contrastive Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 22650–22658.

[49] Qinghai Zheng, Jihua Zhu, and Zhongyu Li. 2021. Collaborative unsupervised multi-view representation learning. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 7 (2021), 4202–4210.