1 Related Work

1.1 Contrastive Learning

Contrastive learning has become a dominant approach for representation learning, aiming to distinguish between positive and negative samples. Owing to its remarkable success, contrastive learning has been extended to clustering tasks, where the objective is not only to learn discriminative representations but also to preserve semantic grouping information. For example, Pan et al. [1] integrated view information and graph contrastive regularization to learn a high-quality consensus graph for final clustering. Zhang et al. [2] mapped samples to a joint space using optimal transport, then took these semantics as supervisory signals to enhance cross-view contrastive learning, thereby improving clustering accuracy. Lu et al. [3] preserved modal-specific information by using cross-modal contrastive learning, achieving higher clustering performance. Chang et al. [4] incorporated sample uncertainty into contrastive learning, dynamically adjusting the weight of contrastive loss to reduce interference from noisy samples.

1.2 Deep Multi-view Clustering

As a central research direction in unsupervised learning, multi-view clustering (MVC) focuses on exploiting the complementary information provided by multiple views of the same object in order to achieve more accurate clustering results. Recent advances in deep learning have motivated its application to multi-view clustering [2,3,5–7]. Deep multi-view clustering leverages deep neural networks to learn nonlinear and hierarchical cross-view representations, further improving clustering performance. For example, (1) Graph neural network-based methods: Zhao et al. [8] leveraged a learnable graph to guide the modeling of robust multi-view representations for clustering tasks. (2) Information bottleneck-based methods: Hu et al. [9] enhanced the feature discriminability by embedding the information bottleneck principle into multi-modal representation learning. (3) Subspace-based methods: Lin et al. [10] integrated Transformer-based autoencoders with low-rank subspace learning and finally performed spectral clustering on the affinity matrix to obtain promising clustering results. (4) Contrastive learning-based methods: Wang et al. [11] performed contrastive learning at three granularities based on the affinity matrix to align cross-view representations.

In recent years, multi-view clustering methods based on contrastive learning have attracted extensive attention. They can be broadly divided into three categories: (1) Feature-level contrastive methods [10, 12, 13]. These approaches perform contrastive learning directly on the instance-level features to maximize the agreement of the same sample across different views. For example, a selective alignment strategy is introduced [13], which aligns representations of different views at the sample level to boost clustering accuracy. (2) Cluster-level contrastive methods [5, 7, 14]. Instead of contrasting raw features, these methods contrast high-level cluster assignments or semantic distributions across views. For example, Chen et al. [14] explored the semantic information by contrasting the clustering assignments across multiple views to ensure consistent label representations. (3) Hybrid methods combining feature-level and cluster-level contrast [15–18]. For example, Hu et al. [17] proposed a joint contrastive triple-learning framework to capture complementary feature and label information across different views. However, the above methods only focus on aligning instances themselves across either feature-level or label-level representations, while neglecting a crucial aspect of the consistency of structural relationships between different instances. In multi-view data, although feature distributions may vary across views, the relational structure among samples is expected to remain consistent. For instance, in an image–text dataset, if two images are visually similar, their corresponding textual descriptions are also likely to exhibit semantic correlation. Nevertheless, many existing approaches fail to explicitly account for such cross-view structural consistency, which may still be semantic confusion under different views, causing intra-class dispersion and inter-class overlap.

Compared with the above categories, our proposed RCMVC introduces relation-level contrastive learning, which aligns cross-view pairwise relationships rather than only individual samples or cluster distributions. We further incorporate a global guidance mechanism that provides feature–semantic guidance for each local view.

Table 1 Nomenclature.

Symbol	Description
X^v	Input data from the v-th view.
x_i^v	Sample i from X^v.
Z^v	Latent feature representation from the v-th view.
z_i^v	Sample i from Z^v.
H^s	Global feature from global view s.
h_s^i	Sample i from H^s.
P^v	Local clustering assignments from the v-th view.
p_i^v	Clustering assignments of the i-th sample from the v-th view.
Q^s	Global semantics from global view s.
q_s^i	Sample i from Q^s.
m	Number of views.
k	Number of clusters.
n	Number of samples in each view.
α, β	Balancing parameters in the objective function.

2 The Proposed Method

2.1 Problem Formulation

Consider that $\{X^1, X^2, \ldots, X^m\}$ denotes m distinct input views. For the i-th view, the observed data is denoted as $X^i = \{x_1^i, x_2^i, ..., x_n^i\} \in \mathbb{R}^{n \times d^i}$, where d^i is the feature dimension of the i-th view and n is the number of samples shared across all views. Each instance $x_j^i \in \mathbb{R}^{d^i}$ represents a feature vector describing the j-th sample in view i. $\{Z^1, Z^2, \ldots, Z^m\}$ denote the latent feature representations extracted from $\{X^1, X^2, \ldots, X^m\}$. H^s represents the global feature information obtained by fusing the multi-view features $\{Z^i\}_{i=1}^m$. Correspondingly, $\{P^1, P^2, \ldots, P^m\}$ denotes the local clustering assignments and

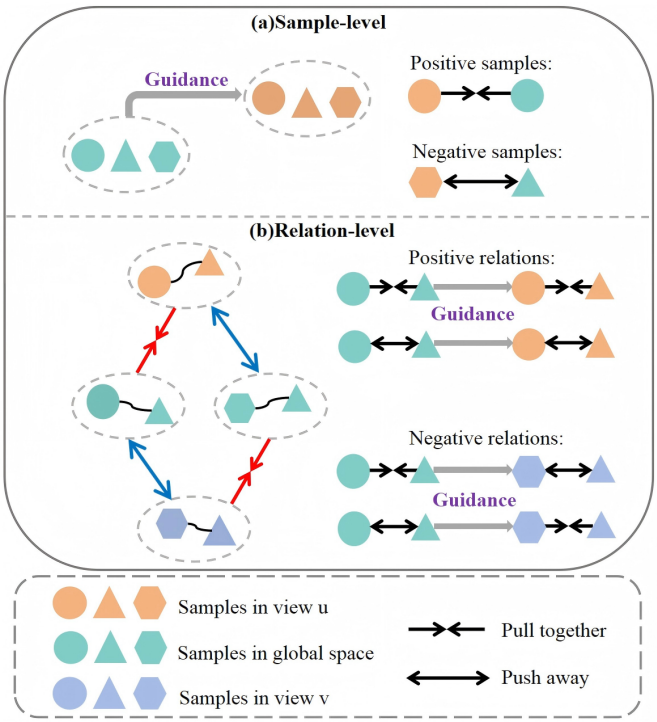

Fig. 1 Contrast process. Positive relations denote the consistent structure of the same sample pairs constructed across views, and negative relations capture the inconsistent structure between heterogeneous sample pairs across such spaces. Accordingly, positive relations are encouraged to be aligned, ensuring that identical samples remain consistently close or distant in both spaces. In contrast, negative relations are enforced to diverge, such that dissimilar samples exhibit inconsistent spatial relations, being close in one space but distant in the other space.

Q^s represents the global semantic information. Note that the number of clusters k is known for the benchmark datasets. This is a widely accepted setting in clustering tasks. In order to enhance clarity, we list the main notations used in Table 1.

2.2 Contrast Process

The details of RCMVC are shown in Fig. 1.

2.3 Global Guidance Mechanism

Local representations are guided by both global feature information and semantic information, yielding representations that are informative, discriminative and robust.

To align local representations with the global feature information, we introduce a contrastive loss that pulls the local representation closer to its corresponding global representation and pushes it away from other samples:

$$\mathcal{L}_{SFC} = -\frac{1}{(m-1)n}\sum_{u=1}^{m}\sum_{i=1}^{n}\log\frac{e^{s(z_i^u,h_i^s)/\tau_1}}{\sum_{j=1}^{n}(e^{s\left(z_i^u,h_j^s\right)/\tau_1})}, \tag{1}$$

where z_i^u denotes sample i from the u-th view, h_i^s denotes sample i from global feature and h_j^s denotes sample j from global feature.

The above loss establishes a contrastive alignment between the local-view representation z_i^u and its corresponding global representation h_i^s. In this way, the global feature representation h_i^s integrates complementary information from all views. Specifically, (z_i^u, h_i^s) is treated as a positive pair and will be pulled closer, while (z_i^u, h_j^s) forms a negative pair and will be pushed away. This objective encourages each local-view embedding z_i^u to capture information consistent with the global view h_i^s and differ from h_j^s, thus mitigating the insufficient information problem in local views.

To incorporate semantic guidance, we use a contrastive objective between local clustering assignments and their corresponding global semantic information:

$$\mathcal{L}_{SLC} = -\frac{1}{k}\sum_{u=1}^{m}\sum_{i=1}^{k}\log\frac{e^{s(p_i^u,q_i^s)/\tau_2}}{\sum_{j=1}^{n}e^{s\left(p_i^u,q_j^s\right)/\tau_2}} - H(Y), \tag{2}$$

where p_i^u denotes clustering assignments of the i-th sample from the u-th view, q_i^s denotes sample i from global semantics and q_j^s denotes sample j from global semantics.

The global semantic information q_i^s encodes cluster information and provides a stable reference for local representation. p_i^u is expected to be close to q_i^s but away from q_j^s. This objective encourages each local cluster assignment p_i^u to capture information consistent with the global semantic information q_i^s. It aligns local representations with their semantic prototypes to ensure that features within the same semantic group are compact, while different groups are separated, improving discriminability.

2.4 Differences with Related Methods

Compared with existing deep multi-view clustering methods, our approach RCMVC differs in the following aspects: (1) While most existing methods rely solely on sample-level contrast, RCMVC introduces a relation-level contrastive module that explicitly aligns pairwise structural relationships across views. This helps the model preserve the underlying data structure and improves robustness to view imbalance. (2) RCMVC further incorporates a global guidance mechanism that utilizes both feature-level and semantic-level information from the global representation space to guide the learning of individual views. This helps mitigate the adverse effects of noisy or biased predictions in specific views and ensures more consistent and reliable cross-view representations. Thus, RCMVC enhances clustering quality by tackling the issues of inconsistent structural alignment and insufficient global guidance.

2.5 Clustering instead of Prediction

In many real-world predictive tasks, obtaining large amounts of high-quality labeled data is extremely costly, requiring significant human effort, expert knowledge, time, and financial resources. Therefore, compared with supervised learning methods that rely on annotated labels, unsupervised clustering provides a more practical and scalable solution, as it can automatically discover the intrinsic structure and semantic information of data without requiring manual annotations. Although some benchmark datasets provide true labels, our model never uses the true labels during training and the true labels are used only for evaluation.

Table 2 Details of five well-known multi-view datasets.

Dataset	# View	# Samples	# Clusters	# Dimensionality
Caltech-2V	2	1400	7	40/254
Caltech-3V	3	1400	7	40/254/928
Event8	3	1579	8	1000/1000/1000
Hdigit	2	10000	10	784/256
NUS22	2	10155	22	1000/1000

2.6 Differences with Multi-modal Clustering

Multi-modal clustering can be regarded as a special case of multi-view clustering. Multi-view clustering focuses on integrating multiple views of the same samples, regardless of whether these views originate from different feature extractors, sensors, or modalities. Multi-modal clustering typically refers to cases where the views correspond to different data modalities (e.g., image–text and audio–visual), which usually exhibit greater heterogeneity. Although multi-modal data require more sophisticated alignment or fusion strategies due to larger semantic and statistical gaps, they still fit naturally into the broader multi-view framework. Thus, multi-modal methods can be seen as multi-view methods specialized for particular heterogeneous views.

3 Experiments

3.1 Datasets

Five well-known multi-view datasets are shown in Table 2.

Caltech-2V [19] contains a total of 1,400 images categorized into 7 classes. **Caltech-3V** [19] extends Caltech-2V by adding one more feature. **Event8** [20] is a challenging image dataset composed of 1,579 samples across eight types of sports events. **Hdigit** [21] contains 10,000 samples with two distint views. **NUS22** [22] is a subset selected from the NUS-WIDE-Object dataset, comprising 10,155 images.

Table 3 Clustering performance on different datasets.

Methods	Caltech-2V		Caltech-3V		Event8		Hdigit		NUS22	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
KM	41.6	30.5	46.3	31.3	34.7	20.7	52.9	47.2	12.5	8.4
Ncuts	39.9	31.2	42.6	25.4	34.8	15.5	60.0	45.9	12.9	7.4
ALLKM	46.4	31.4	46.9	31.5	28.7	11.6	76.1	74.8	12.8	7.1
AllNcuts	42.8	25.2	43.7	25.5	35.2	20.3	67.5	58.7	14.9	9.4
RMKMC	51.4	33.5	59.5	49.4	37.8	27.3	79.8	80.1	14.3	9.9
LMVSC	56.0	43.2	70.4	58.9	42.2	25.9	97.0	92.9	15.5	12.9
FPMVS-CAG	55.1	41.9	68.7	53.2	48.9	32.0	91.0	81.6	17.2	12.3
RCMVC	**69.3**	**56.4**	**77.7**	**64.1**	**59.0**	**40.7**	**99.8**	**99.4**	**21.2**	**14.2**

3.2 State-of-the-art Methods

We compared the proposed RCMVC with four kinds of state-of-the-art clustering methods.

Single-view Clustering Methods: K-Means (KM), Normalized Cuts (Ncuts).

All-view Clustering Methods: AllKM, AllNcuts, implemented by applying the above single-view clustering methods to the concatenated multi-view features.

Traditional Multi-view Clustering Methods: RMKMC [23], LMVSC [24], FPMVS-CAG [25].

Deep Multi-view Clustering Methods: MvSCN [1] [26], EAMC [27], DEMVC [2] [28], SiMVC and CoMVC [3] [13], MFLVC [4] [29], SPDMC [5] [30], CVCL [6] [14], ICMVC [7] [31], DIVIDE [8] [5], PDMC-RCL [9] [15] and MSDIB [10] [9].

3.3 Implementation Details

We implemented RCMVC using PyTorch 1.12.0 (Python 3.9) on a Windows 10 platform equipped with an NVIDIA RTX 4090D GPU (24 GB) and an Intel i7-14700K CPU. In the experiments, we observed that 100 epochs were enough to achieve stable performance. Each experiment was repeated ten times and the clustering result with the lowest loss was reported. Our method does not involve any data splitting. Since clustering is a fully unsupervised task, all samples are used directly for representation learning and clustering. Therefore, there is no need for data splits, and the training process is conducted in a batch-based manner, following common practice in deep multi-view clustering tasks. The training was performed with a batch size of 256 using the Adam optimizer with a learning rate of 0.0003. The temperature hyperparameters τ_1, τ_2, τ_3, and τ_4 in the four loss functions are all set to 1.0. The trade-off hyperparameters α and β were tuned via grid search in the range (0, 1) with a step size of 0.2. To evaluate the clustering performance, we adopted two widely used clustering metrics: Clustering Accuracy (ACC) and Normalized Mutual Information (NMI), where higher scores imply better clustering results.

3.4 Experimental Results

From the experimental results shown in the Table 3, our method achieves consistent and substantial improvements over both single-view and all-view clustering methods across all datasets. This demonstrates that RCMVC is able to effectively exploit complementary information from multiple views, going far beyond simple feature concatenation or single-view representation learning. Benefiting from expressive power of deep learning frameworks, most deep multi-view clustering methods outperform traditional clustering methods.

3.5 Ablation Study

In this section, the effectiveness of each component of RCMVC is further verified by ablation studies. The results shown in Table 4 demonstrate that the model yields the poorest performance when including only $\mathcal{L}_{REC}$ in the loss function. However, the incorporation of $\mathcal{L}_{SCL}$ or $\mathcal{L}_{RCL}$ significantly improves the clustering performance across all datasets. Adding only $\mathcal{L}_{SCL}$ leads to a more significant performance improvement, demonstrating that enhancing sample-level discriminability is crucial for representation quality. Adding only $\mathcal{L}_{RCL}$ also improves performance, but to a lesser extent, indicating that relation-level contrast helps but is insufficient on its own. Because learning discriminative representations is of primary importance, as it enables effective

[1] Website at github.com/XLearning-SCU/2019-IJCAI-MvSCN

[2] Website at github.com/SubmissionsIn/DEMVC

[3] Website at github.com/DanielTrosten/mvc

[4] Website at github.com/SubmissionsIn/MFLVC

[5] Website at www.researchgate.net/publication/376311276_2023-TNNLS-SPDMC-demo

[6] Website at github.com/chenjie20/CVCL

[7] Website at github.com/liunian-Jay/ICMVC

[8] Website at github.com/XLearning-SCU/2024-AAAI-DIVIDE

[9] Website at github.com/ShizheHu/TIP25_Code_PDMC-RCL

[10] Website at github.com/ShizheHu/AAAI25_Code_MSDIB

Table 4 Ablation study on different datasets.

Methods	Caltech-2V		Caltech-3V		Event8		Hdigit		NUS22	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
(1) $\mathcal{L}_{REC}$	52.8	42.2	57.5	47.3	28.5	13.4	45.9	31.9	11.7	5.7
(2) $\mathcal{L}_{REC} + \mathcal{L}_{SCL}(\beta = 0)$	65.3	52.2	59.0	54.6	50.7	36.6	91.7	83.2	17.6	12.8
(3) $\mathcal{L}_{REC} + \mathcal{L}_{RCL}(\alpha = 0)$	54.4	36.6	58.2	41.5	39.3	20.0	69.4	51.1	13.5	7.1
(4) $\mathcal{L}_{REC} + \mathcal{L}_{SCL} + \mathcal{L}_{RCL}$	**69.3**	**56.4**	**77.7**	**64.1**	**59.0**	**40.7**	**99.8**	**99.4**	**21.2**	**14.2**

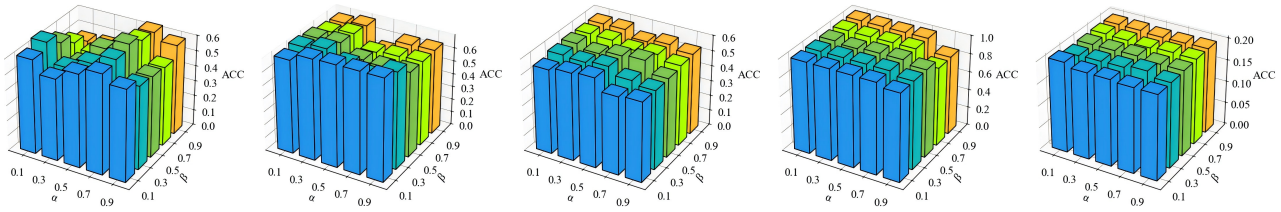

Fig. 2 Parameter analysis of RCMVC on Caltech-2V, Caltech-3V, Event8, Hdigit and NUS22 datasets.

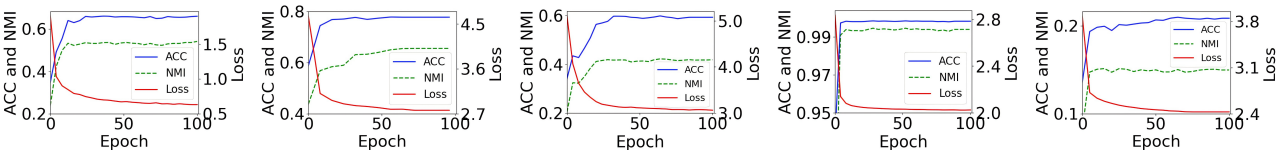

Fig. 3 Convergence analysis of RCMVC on Caltech-2V, Caltech-3V, Event8, Hdigit and NUS22 datasets.

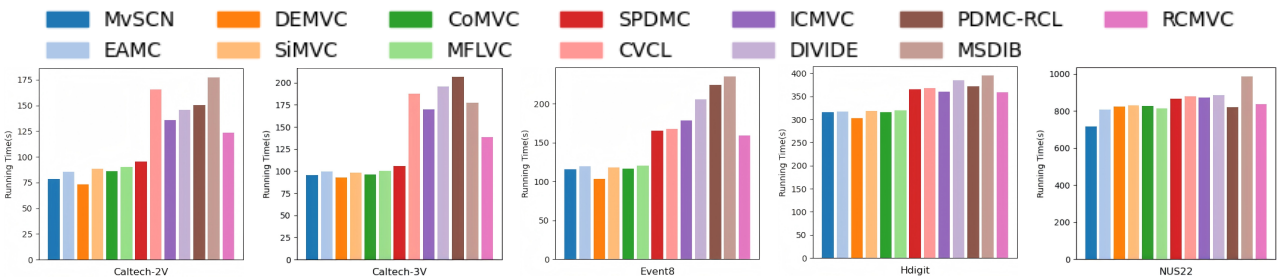

Fig. 4 Running time analysis of RCMVC on Caltech-2V, Caltech-3V, Event8, Hdigit and NUS22 datasets.

separation of different categories. On this basis, incorporating cross-sample relational consistency can further enhance structural alignment and semantic coherence across views. The full model using both losses achieves the best performance, confirming that the two components are complementary. When all three loss terms are integrated, our method achieves the best performance. These findings indicate that each component contributes positively to the final clustering results, validating the necessity of their joint optimization.

3.6 Parameter Analysis

To assess how our method responds to different trade-off parameter configurations, we performed a grid search over α and β, setting their values from 0 to 1 with a step size of 0.2. The results, depicted in Fig. 2, show that clustering performance remains largely stable across most datasets, indicating that the proposed method is relatively insensitive to different parameters. For the Caltech-2V and Caltech-3V datasets, we notice fluctuations in ACC, which indicates that the influence of sample-level and relation-level contrastive losses is not uniform across datasets. Hence, clustering tasks with different data characteristics may require distinct balances of cross-view alignment and consistency.

3.7 Convergence Analysis

To verify the convergence of RCMVC, we plotted the variation of the total loss over training epochs in Fig. 3. As shown, the loss decreases rapidly during the early training phase and gradually converges to a fixed value. In addition, we also visualize the clustering performance in terms of ACC and NMI. Both metrics steadily increase as training progresses, eventually reaching optimal values. These results confirm that our method converges effectively.

3.8 Running Time Analysis

To further evaluate the efficiency of the proposed framework, we compare the computational cost of RCMVC with several representative deep multi-view clustering baselines. We run all the models for 100 epochs, and the results are shown in Fig. 4. The results show that RCMVC incurs a slightly higher training cost than lightweight contrastive frameworks such as CoMVC and CVCL, mainly due to the introduction of relation-level contrastive learning and the global-guidance

mechanism. However, it is worthy because RCMVC can achieve satisfactory clustering results.

References

[1] Pan E, Kang Z. Multi-view contrastive graph clustering. Advances in neural information processing systems, 2021, 34: 2148–2159

[2] Zhang Q, Zhang L, Song R, Cong R, Liu Y, Zhang W. Learning common semantics via optimal transport for contrastive multi-view clustering. IEEE Transactions on Image Processing, 2024

[3] Lu Y, Li Q, Zhang X, Gao Q. Deep contrastive representation learning for multi-modal clustering. Neurocomputing, 2024, 581: 127523

[4] Chang L, Chen L, Zhou C. Uncertainty-aware contrastive learning for deep clustering. Neurocomputing, 2025, 130568

[5] Lu Y, Lin Y, Yang M, Peng D, Hu P, Peng X. Decoupled contrastive multi-view clustering with high-order random walks. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(13): 14193–14201

[6] Fei L, He J, Zhu Q, Zhao S, Wen J, Xu Y. Deep multi-view contrastive clustering via graph structure awareness. IEEE Transactions on Image Processing, 2025

[7] Duan Y, Chen H, Zhang R, Wang R, Nie F, Li X. Soft neighbors supported contrastive clustering. IEEE Transactions on Image Processing, 2025

[8] Zhao L, Wang X, Liu Z, Wang Z, Chen Z. Learnable graph guided deep multi-view representation learning via information bottleneck. IEEE Transactions on Circuits and Systems for Video Technology, 2024

[9] Hu S, Fan J, Zou G, Ye Y. Multi-aspect self-guided deep information bottleneck for multi-modal clustering. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(16): 17314–17322

[10] Lin Y, Liu H, Yu X, Zhang C. Leveraging transformer-based autoencoders for low-rank multi-view subspace clustering. Pattern Recognition, 2025, 161: 111331

[11] Wang J, Feng S, Lyu G, Gu Z. Triple-granularity contrastive learning for deep multi-view subspace clustering. In: Proceedings of the 31st ACM international conference on multimedia. 2023, 2994–3002

[12] Wu S, Zheng Y, Ren Y, He J, Pu X, Huang S, Hao Z, He L. Self-weighted contrastive fusion for deep multi-view clustering. IEEE Transactions on Multimedia, 2024, 26: 9150–9162

[13] Trosten D J, Løkse S, Jenssen R, Kampffmeyer M. Reconsidering representation alignment for multi-view clustering. In: IEEE Conference on Computer Vision and Pattern Recognition. 2021, 1255–1265

[14] Chen J, Mao H, Woo W L, Peng X. Deep multiview clustering by contrasting cluster assignments. In: 2023 IEEE/CVF International Conference on Computer Vision. 2023, 16706–16715

[15] Lou Z, Xue H, Wang Y, Zhang C, Yang X, Hu S. Parameter-free deep multi-modal clustering with reliable contrastive learning. IEEE Transactions on Image Processing, 2025, 34: 2628–2640

[16] Xu J, Meng M, Liu J, Wu J. Deep multi-view clustering with diverse and discriminative feature learning. Pattern Recognition, 2025, 161: 111322

[17] Hu S, Zou G, Zhang C, Lou Z, Geng R, Ye Y. Joint contrastive triple-learning for deep multi-view clustering. Information Processing & Management, 2023, 60(3): 103284

[18] Shu Z, Sun T, Luo Y, Yu Z. Ambiguous instance-aware contrastive network with multi-level matching for multi-view document clustering. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(19): 20479–20487

[19] Fei-Fei L, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: 2004 conference on computer vision and pattern recognition workshop. 2004, 178–178

[20] Li L J, Fei-Fei L. What, where and who? classifying events by scene and object recognition. In: 2007 IEEE 11th International Conference on Computer Vision. 2007, 1–8

[21] Chen M, Lin J, Li X, Liu B, Wang C, Huang D, Lai J. Representation learning in multi-view clustering: A literature review. Data Science and Engineering, 2022, 7(3): 225–241

[22] Chua T S, Tang J, Hong R, Li H, Luo Z, Zheng Y. 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. 2009, 1–9

[23] Cai X, Nie F, Huang H. Multi-view k-means clustering on big data. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. 2013, 2598–2604

[24] Kang Z, Zhou W, Zhao Z, Shao J, Han M, Xu Z. Large-scale multi-view subspace clustering in linear time. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(04): 4412–4419

[25] Wang S, Liu X, Zhu X, Zhang P, Zhang Y, Gao F, Zhu E. Fast parameter-free multi-view subspace clustering with consensus anchor guidance. IEEE Transactions on Image Processing, 2022, 31: 556–568

[26] Huang Z, Zhou J T, Peng X, Zhang C, Zhu H, Lv J. Multi-view spectral clustering network. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. 2019, 2563–2569

[27] Zhou R, Shen Y D. End-to-end adversarial-attention network for multi-modal clustering. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 14607–14616

[28] Xu J, Ren Y, Li G, Pan L, Zhu C, Xu Z. Deep embedded multi-view clustering with collaborative training. Information Sciences, 2021, 573: 279–290

[29] Xu J, Tang H, Ren Y, Peng L, Zhu X, He L. Multi-level feature learning for contrastive multi-view clustering. In: IEEE Conference on Computer Vision and Pattern Recognition. 2022, 16030–16039

[30] Chen R, Tang Y, Xie Y, Feng W, Zhang W. Semisupervised progressive representation learning for deep multiview clustering. IEEE Transactions on Neural Networks and Learning Systems, 2023, 35(10): 14341–14355

[31] Chao G, Jiang Y, Chu D. Incomplete contrastive multi-view clustering with high-confidence guiding. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 11221–11229