Relational Contrastive Learning For Multi-View Clustering

1st Zhonghua Li
Computer Science of South China Normal University
Guangzhou, China
Atonioy@163.com

2nd Zhehan Zhou
Computer Science of South China Normal University
Guangzhou, China
leslielalala12@gmail.com

3rd Na Tang
Computer Science of South China Normal University
Guangzhou, China
tangna@scnu.edu.cn

4th Xiaolin Xiao*
Computer Science of South China Normal University
Guangzhou, China
shellyxiaolin@gmail.com*

***Abstract*—Multi-View Clustering (MVC) has recently benefited from contrastive learning, which enhances representation discriminability by pulling positive pairs together and pushing negative pairs apart. However, existing contrastive MVC methods fail to preserve the consistency of pairwise similarities across samples. They either treat different views of the same instance as positives, overlooking intra-cluster relations, or rely on pseudo-labels that uniformly assign positives and negatives, neglecting the pairwise similarities. To address these issues, we propose a relational contrastive learning for multi-view clustering (RCMVC) framework that models the similarity distribution between an anchor and all other samples, explicitly preserving pairwise consistency. This design enables soft positive and negative assignments, capturing fine-grained similarities while maintaining coherent cluster structures. Moreover, RCMVC incorporates both instance-level and cluster-level modules to jointly exploit detailed and semantic relations. Extensive experiments demonstrate the importance of preserving the consistency of pairwise similarities for multi-view clustering.**

***Index Terms*—Multi-view clustering, contrastive learning, relational learning**

I. Introduction

Multi-View Clustering (MVC) integrates data from multiple sources or perspectives to reveal intrinsic structures and improve clustering performance. Among existing approaches, neural network-based methods [1], [2] enhance representation learning via deep architectures, yet often lack sufficient discriminative power in unsupervised settings. Recently, contrastive learning, which constructs positive and negative pairs to improve feature discriminability [3], has been integrated into MVC, enhancing cross-view consistency and boosting clustering performance.

Although good empirical performance has been observed [4]–[6], existing contrastive MVC methods still face notable limitations. Many approaches tend to treat all samples with the same pseudo-label as equally "positive" and all samples with different pseudo-labels as equally "negative", completely ignoring the consistency of pairwise similarities among different samples. Based on this observation, we propose introducing relational learning to accurately capture the underlying relationships among samples. As illustrated in Fig. 1 (a), assuming that samples S_1 and S_2 share the same pseudo-label as the anchor and samples S_3 and S_4 exhibit different pseudo-labels as the anchor. Previous methods uniformly pulled S_1 and S_2 equally close to the anchor and pushed S_3 and S_4 equally far from the anchor. These approaches capture semantic consistency but are too coarse and fail to capture instance-level relationships. In contrast, we outline our basic idea in Fig. 1 (b). Specifically, if S_1 shows higher similarity to the anchor, this relationship should be preserved, and S_1 should be pulled closer to the anchor compared to S_2. Similarly, if S_3 shows lower similarity to the anchor, this relative relationship should be maintained

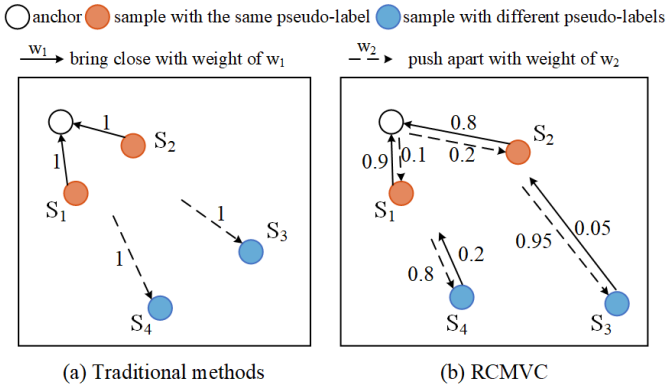


Fig. 1: Key differences between traditional methods and our RCMVC framework: (a) Given S_1 and S_2 that share the same pseudo-label as the anchor, previous methods uniformly pulled S_1 and S_2 equally close to the anchor and pushed S_3 and S_4 equally far away. This means that all positive pairs were treated as having the same degree of semantic relevance, and all negative pairs were penalized equally regardless of their actual similarity. As a result, the fine-grained relationships among samples were often ignored in the representation learning process. (b) In RCMVC, if S_1 has a higher similarity to the anchor than S_2, this relationship is preserved by pulling S_1 closer than S_2. Similarly, for negative samples S_3 and S_4 with different pseudo-labels from the anchor, if S_3 is less similar to the anchor than S_4, it is pushed farther away compared to S_4.

by pushing S_3 farther away from the anchor compared to S_4. In this way, the consistency of pairwise similarities across all samples is preserved by considering instance-level relationships, rather than just cluster-level semantics. Moreover, pseudo-labels are highly inaccurate in unsupervised settings. Methods that directly pull together samples with the same pseudo-label and push apart samples with different pseudo-labels heavily depend on the accuracy of these pseudo-labels. Our model alleviates this issue by considering the instance-level relationships, enabling a better capture of the intrinsic data structure.

To summarize, we propose a relational contrastive learning for multi-view clustering (RCMVC) framework. Specifically, we first use an autoencoder to learn low-level features from individual views and then stack two MLPs to obtain high-level features shared across all views. Following this, we design an instance-level relational contrastive learning module, where the similarity distribution between the anchor and all other samples is computed, and the consistency of these pairwise similarities is preserved during representation learning. In other words, instead of rigidly classifying a sample as either positive or negative, we introduce relational learning to enable a soft assignment of positive and negative samples. Meanwhile, following the idea of labels as representations [7], we design a cluster-level relational contrastive learning module to account for semantics consistency. Compared to previous work, our contributions are as follows:

- We propose a relational contrastive learning framework for multi-view clustering. Our model allows for a flexible assignment of samples by preserving the consistency of pairwise similarities, capturing complex data structures and potentially mitigating the side effects of inaccurate pseudo-labels.
- We implement relational contrastive learning at both instance and cluster levels, jointly capturing fine-grained data relationships and the semantic consistency.
- In the experiments, we demonstrate the superiority and effectiveness of our RCMVC model compared to state-of-the-art methods. Additionally, the effectiveness of relational contrastive learning is evaluated through an ablation study.

II. Related Work

A. Contrastive Learning

Contrastive learning [8] has achieved considerable success in the field of self-supervised learning. Contrastive learning is able to learn more efficient and robust feature representations by pulling similar pairs closer and pushing dissimilar pairs apart. Momentum Contrast for Unsupervised Visual Representation Learning (MoCo) [9] proposes contrast learning as the perspective of dictionary search, and constructs a dynamic dictionary with queue and moving average encoder. This makes it possible to build a large and consistent dictionary in real time, facilitating contrast unsupervised learning. Contrastive clustering (CC) [7] is a single stage online clustering method, which simultaneously optimize instance and cluster-level contrastive loss, the model combines learning representation and cluster assignment in an end-to-end manner. Contrastive learning with stronger augmentations (CLSA) [10] proposes a general contrastive learning framework to complement the existing contrast learning methods. It utilizes the distribution differences between weakly and strongly enhanced images on the representation library to oversee the retrieval of strongly enhanced queries from the pool of instances. Prototypical contrastive learning of unsupervised representations (PCL) [11] not only learns low-level features for instance discriminant tasks, but more importantly encodes semantic structures discovered through clustering into the learned embeddings space.

B. Deep Multi-view Representation Learning

In recent years, deep Multi-view representation Learning has become increasingly popular among a wide range of researchers because of its outstanding representation capabilities. Deep embedded multi-view clustering with collaborative training (DEMVC) [2] proposes to train multiple deep neural networks collaboratively, which can make good use of the common and complementary information of multiple views. In order to improve the performance of MVC, a sharing scheme of auxiliary distribution and a new cluster center initialization consistency strategy are proposed. Deep embedding clustering based on contractive autoencoder (DECCA) [12] retains important information from the original data, and combines the original sample with its corresponding enhanced representation constraints to achieve representation learning. In addition, in order to ensure the local properties of the class clusters, the constraint of neighbor sample similarity is introduced.

Following a phase of rapid advancement, deep multi-view representation learning has faced a performance plateau, primarily due to its limited discriminative capability in unsupervised settings. Recently, contrastive learning has emerged as a powerful approach in unsupervised learning, boosting discriminative performance by leveraging contrasts between positive and negative pairs. Multi-level Feature Learning for Contrastive Multi-view Clustering (MFLVC) [13] designs a non-fusion MVC model to avoid merging unfavorable view private information between all views, and generates different levels of features for the samples in each view, including low-level features, high-level features and semantic features. Global and cross-view feature aggregation for multi-view clustering (GCFAgg) [14] aligns the consistent representation with the view-specific representation through a structure-guided contrastive learning module so that view-specific representations with high structural relationships from different samples are similar.

III. PROPOSED METHOD

In this section, we present relational contrastive learning for multi-view clustering. As illustrated in Fig. 2, the framework consists of three modules.

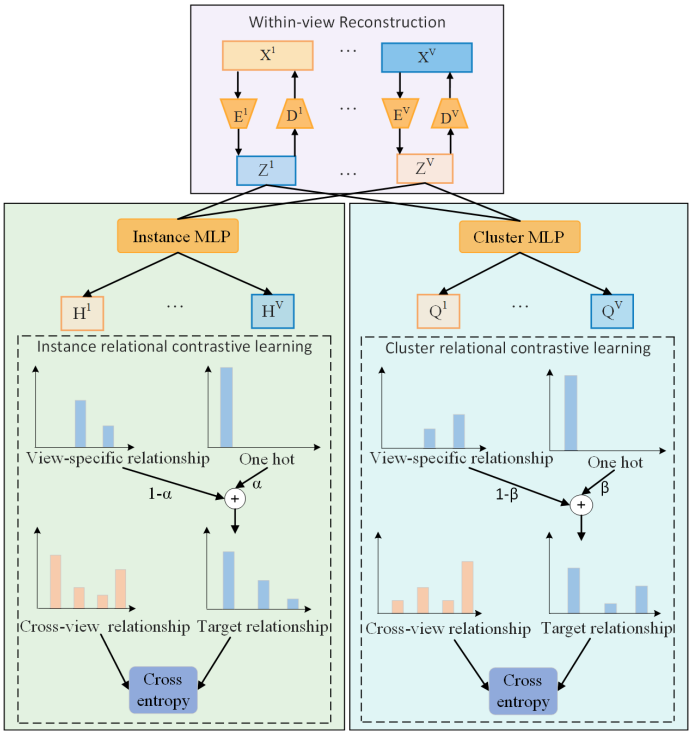


Fig. 2: The framework of RCMVC. Encoder $\{E^v\}_{v=1}^V$, decoder $\{D^v\}_{v=1}^V$, view-specific representation Z^v, instance-level representation H^v, cluster-level representation Q^v.

A. Dataset Description

Given a multi-view dataset $\left\{X^v \in \mathbb{R}^{N\times D_v}\right\}_{v=1}^V$ with N samples from V views and K categories, where $X^v = [x_1^v, x_2^v, \ldots, x_{\mathrm{N}}^v]$ denotes the instances in the v_{th} view, and D_v represents the dimension of the v_{th} view.

B. Within-view Reconstruction

To mitigate noise and enhance feature quality, we introduce autoencoders as view-specific feature extractors, which are widely used in unsupervised tasks [15]. For the v_{th} view, we define an encoder $f_{\theta^v}^v$ and a decoder $g_{\eta^v}^v$, where θ^v and η^v are model parameters. The encoder maps an input sample into a latent feature space:

$$\mathbf{z}_i^v = f_{\theta^v}^v\left(\mathbf{x}_i^v\right), \tag{1}$$

where $z_i^v \in \mathbb{R}^d$ is the latent representation of the i_{th} sample from the v_{th} view. The decoder reconstructs the input as:

$$\hat{\mathbf{x}}_i^v = g_{\eta^v}^v(\mathbf{z}_i^v). \tag{2}$$

The reconstruction loss is minimized across all views:

$$\mathcal{L}_{\mathrm{Re}} = \sum_{v=1}^{V}\sum_{i=1}^{N}\left\|\mathrm{x}_i^v - g_{\eta^v}^v\left(f_{\theta^v}^v\left(\mathbf{x}_i^v\right)\right)\right\|_2^2. \tag{3}$$

C. Instance-level Relational Contrastive Learning

Traditional contrastive MVC methods usually treat different views of the same sample as positive pairs, neglecting the high similarity between samples within the same cluster. To better capture complex data relationships, we introduce relational contrastive learning to preserve the consistency of pairwise similarities.

While autoencoders minimize reconstruction error, they tend to emphasize view-specific features over cross-view consistency. To address this, we apply a shared MLP, parameterized by G_H, to the view-specific features $\{Z^v\}_{v=1}^V$ to obtain high-level embeddings $\{H^v\}_{v=1}^V$. The cosine similarity between embeddings is computed as:

$$sim(\mathbf{h}_i^v, \mathbf{h}_j^n) = \frac{\langle \mathbf{h}_i^v, \mathbf{h}_j^n \rangle}{\|\mathbf{h}_i^v\|_2 \|\mathbf{h}_j^n\|_2}, \tag{4}$$

where $\langle , \rangle$ denotes the dot product.

Given an anchor h_i^v, we compute its similarity with all other samples in the view, and then apply a softmax layer to obtain a similarity distribution $R_i \in \mathbb{R}^N$. The p_{th} element of R_i is computed as:

$$r_{ip} = \frac{\mathbb{1}_{p\neq i} e^{sim(\mathbf{h}_i^v, \mathbf{h}_p^v)/\tau_1}}{\sum_{j=1}^N \mathbb{1}_{j\neq i} e^{sim(\mathbf{h}_i^v, \mathbf{h}_j^v)/\tau_1}}, \tag{5}$$

where $\mathbb{1}$ is the indicator function and τ_1 is the temperature parameter. Traditional contrastive MVC assumes positives come only from other views of the anchor, while overlooking potential positive relationships within the same view. Instead, we define a target similarity distribution $W_i \in \mathbb{R}^N$, whose p_{th} element is computed as:

$$w_{ip} = \alpha \mathbb{1}_{i=p} + (1-\alpha) r_{ip}, \tag{6}$$

where α is a weighting coefficient. A large w_{ip} indicates a higher probability that the i_{th} sample and the p_{th} sample form a positive pair.

For cross-view similarity, we compute a distribution $T_i \in \mathbb{R}^N$, whose p_{th} element is computed as:

$$t_{ip} = \frac{e^{sim(\mathbf{h}_i^v, \mathbf{h}_p^n)/\tau_1}}{\sum_{j=1}^{N} \sum_{m=v,n} e^{sim(\mathbf{h}_i^v, \mathbf{h}_j^m)/\tau_1} - e^{1/\tau_1}}. \tag{7}$$

We then minimize the cross-entropy between W_i and T_i, defining the relational contrastive loss for anchor i:

$$\ell_{Ir}^{(vn)} = -\frac{1}{N} \sum_{i=1}^{N} W_i \log T_i. \tag{8}$$

Aggregating over all views gives the instance-level relational contrastive loss:

$$\mathcal{L}_{\text{Ins}} = \frac{1}{2} \sum_{v=1}^{V} \sum_{n \neq v} \ell_{Ir}^{(vn)}. \tag{9}$$

D. Cluster-Level Relational Contrastive Learning

Cluster-level structures are crucial for effective clustering. Following the "labels as representations" principle [7], we employ a shared cluster MLP, parameterized by G_C, on the view-specific features $\{Z^v\}_{v=1}^V$ to obtain cluster-level embeddings $\{Q^v \in \mathbb{R}^{N \times K}\}_{v=1}^V$. Each row $Q_{i\cdot}^v \in \mathbb{R}^K$ represents the soft label of the i_{th} instance, encoding its assignment probabilities over the K clusters, while each column $Q_{\cdot j}^v \in \mathbb{R}^N$ can be regarded as the representation of the j_{th} cluster, reflecting its distribution across all instances.

At the cluster level, embeddings of the same cluster across different views form positive pairs, while embeddings of different clusters serve as negative pairs. Specifically, $(Q_{\cdot j}^v, Q_{\cdot j}^n)^{\{v \neq n\}}$ form positive pairs, and $(Q_{\cdot j}^v, Q_{\cdot k}^m)_{\{j \neq k\}}^{\{m=v,n\}}$ form negative pairs.

To preserve the consistency of pairwise similarity at the cluster level, we also adopt a relational contrastive learning strategy. The cosine similarity is computed between cluster-level embeddings, followed by a softmax layer to obtain a similarity distribution $R_j' = \left[r_{j1}', r_{j2}', \ldots, r_{jK}'\right]$. The p_{th} element of R_j' is computed as:

$$r_{jp}' = \frac{\mathbb{1}_{p \neq j} e^{sim(\mathbf{Q}_{\cdot j}^v, \mathbf{Q}_{\cdot p}^v)/\tau_2}}{\sum_{k=1}^{K} \mathbb{1}_{k \neq j} e^{sim(\mathbf{Q}_{\cdot j}^v, \mathbf{Q}_{\cdot k}^v)/\tau_2}}, \tag{10}$$

where τ_2 is the temperature coefficient. We use R_j' to formulate a target similarity distribution $W_j' \in \mathbb{R}^K$, whose p_{th} element is constructed as:

$$w_{jp}' = \beta \mathbb{1}_{j=p} + (1-\beta) r_{jp}', \tag{11}$$

with β as a weighting coefficient. Then, we calculate the similarity between $\mathbf{Q}_{\cdot j}^v$ and all other clusters across all views and apply the softmax layer to the computed similarities to formulate a cross-view similarity distribution $T_j' \in \mathbb{R}^K$, whose p_{th} element is computed as:

$$t_{jp}' = \frac{e^{sim(\mathbf{Q}_{\cdot j}^v, \mathbf{Q}_{\cdot p}^n)/\tau_2}}{\sum_{k=1}^{K} \sum_{m=v,n} e^{sim(\mathbf{Q}_{\cdot j}^v, \mathbf{Q}_{\cdot k}^m)/\tau_2} - e^{1/\tau_2}}. \tag{12}$$

Finally, the cluster-level relational contrastive loss is computed via cross-entropy between W_j' and T_j':

$$\ell_{Cr}^{(vn)} = -\frac{1}{K} \sum_{j=1}^{K} W_j' \log T_j'. \tag{13}$$

and the cumulative cross-view loss is:

$$\mathcal{L}_{\mathbf{Clu}} = \frac{1}{2} \sum_{v=1}^{V} \sum_{n \neq v} \ell_{Cr}^{(vn)}. \tag{14}$$

E. Optimization

The overall objective of the framework combines reconstruction, instance-level, and cluster-level losses:

$$\mathcal{L} = \mathcal{L}_{\text{Re}} + \mathcal{L}_{\text{Ins}} + \mathcal{L}_{\text{Clu}}. \tag{15}$$

The cluster assignment for the i_{th} sample is obtained by averaging the cluster-level embeddings across views and selecting the cluster with the highest probability:

$$y_i = \underset{j}{\text{argmax}} \left(\frac{1}{V} \sum_{v=1}^{V} Q_{i\cdot}^v \right), \tag{16}$$

and the overall clustering result is Y=[y_1, ... ,y_N]. The complete optimization procedure is summarized in Algorithm 1.

Algorithm 1 Relational Contrastive learning for Multi-View Clustering (RCMVC)

Input: Multi-view dataset $\{X^v\}_{v=1}^V$; Number of clusters K; Temperature parameters τ_1 and τ_2; weighting coefficients α and β;

1: Initialize $\{\theta^v, \eta^v\}_{v=1}^{\text{V}}$ by minimizing Eq. (3);
2: Calculate similarity distribution by Eq. (5), Eq. (7), Eq. (10) and Eq. (12);
3: Update G_H, G_C, $\{\theta^v, \eta^v\}_{v=1}^{\text{V}}$ by minimizing Eq. (15);
4: The cluster assignment is determined by Eq. (16);

Output: The label predictions Y=[y_1, ... ,y_N].

IV. EXPERIMENTS

A. Experimental Settings

We evaluate RCMVC on six multi-view datasets. NGs [16] contains 500 news documents, preprocessed into three views and categorized into five classes. Prokaryotic [17] includes 551 genome sequences represented in three views. BBC [18] consists of 685 news articles from three views, grouped into ten categories. Caltech-5V [19] is an RGB image dataset represented by five views with 1,400 images in seven categories, yielding three subsets: Caltech-2V, Caltech-3V, and Caltech-4V. Scene15 [20] contains 4,485 images from 15 scene categories. Hdigit [21] is a handwritten digit dataset with 10,000 images from ten classes.

RCMVC is compared against nine state-of-the-art methods: CONAN [22], EAMC [23], SIMVC [24], COMVC [24], MFLVC [13], CVCL [1], DEMVC [2], FMCSC [25], and SCM [26]. Clustering performance is evaluated using three

TABLE I: Clustering performance of RCMVC with competing methods (Prokaryotic, Hdigit, NGs, BBC)

Method	Prokaryotic			Hdigit			NGs			BBC		
	ACC	**NMI**	**PUR**	**ACC**	**NMI**	**PUR**	**ACC**	**NMI**	**PUR**	**ACC**	**NMI**	**PUR**
CONAN	44.28	10.17	52.81	95.62	91.93	95.62	37.80	12.75	40.60	30.22	11.52	54.74
EAMC	32.30	3.91	56.80	48.78	51.51	49.39	33.40	10.14	34.20	26.57	15.58	33.87
SIMVC	42.28	17.24	67.69	78.54	67.05	78.54	42.20	18.23	45.60	36.20	8.38	41.89
COMVC	41.01	19.47	64.24	90.32	87.13	90.32	41.80	15.05	41.80	42.04	13.91	46.42
MFLVC	48.28	24.90	60.44	94.42	87.50	94.40	88.60	74.68	88.60	62.34	56.65	77.08
CVCL	52.40	25.17	64.60	32.16	24.07	32.27	71.80	49.00	45.98	54.00	30.73	59.50
DEMVC	52.45	25.79	69.69	37.38	32.55	48.16	43.00	27.65	27.93	57.37	32.32	32.32
FMCSC	47.18	20.53	47.37	83.02	67.24	87.43	64.40	53.66	64.40	57.69	27.49	56.83
SCM	47.19	17.95	66.79	95.92	89.81	95.92	95.40	87.63	95.40	60.73	39.55	60.73
RCMVC(Ours)	**56.44**	**28.80**	**69.93**	**97.08**	**92.36**	**97.08**	**96.07**	**88.15**	**96.07**	**67.28**	**65.98**	**80.16**

TABLE II: Clustering performance of RCMVC with competing methods (Caltech Datasets)

Method	Caltech-2V			Caltech-3V			Caltech-4V			Caltech-5V		
	ACC	**NMI**	**PUR**	**ACC**	**NMI**	**PUR**	**ACC**	**NMI**	**PUR**	**ACC**	**NMI**	**PUR**
CONAN	57.50	45.16	57.57	59.14	49.81	59.14	55.71	50.61	57.35	72.07	64.18	72.21
EAMC	49.93	44.49	52.07	50.14	36.17	50.29	47.86	37.09	47.86	39.36	35.40	40.00
SIMVC	50.83	47.15	55.73	56.92	49.53	59.12	61.93	53.62	63.03	71.93	67.71	72.92
COMVC	46.63	42.62	52.72	54.13	50.43	58.42	56.83	56.92	64.63	70.03	68.71	74.62
MFLVC	60.60	52.80	61.60	63.12	56.63	63.92	73.32	65.23	73.42	80.42	70.32	80.43
CVCL	64.79	55.03	64.79	66.64	54.93	66.64	66.43	59.34	68.64	74.57	65.49	59.03
DEMVC	49.86	38.45	52.07	53.36	41.36	53.36	49.29	45.04	51.86	46.00	36.66	49.50
FMCSC	52.78	45.72	50.23	66.14	52.27	68.47	71.50	54.09	73.82	74.10	58.10	74.58
SCM	63.07	48.76	63.29	64.29	48.35	64.36	74.64	60.14	74.64	74.29	68.84	78.79
RCMVC(Ours)	**66.63**	**57.62**	**66.63**	**68.78**	**61.27**	**68.78**	**89.91**	**81.66**	**89.91**	**86.12**	**78.94**	**86.12**

TABLE III: Ablation studies on competing datasets (Prokaryotic, Hdigit, NGs, Scene15)

Components		Prokaryotic			Hdigit			NGs			Scene15		
R_{ins}	R_{clu}	**ACC**	**NMI**	**PUR**	**ACC**	**NMI**	**PUR**	**ACC**	**NMI**	**PUR**	**ACC**	**NMI**	**PUR**
✗	✗	45.92	21.53	57.53	93.41	85.39	93.41	85.80	70.96	85.80	33.16	36.43	37.65
✓	✗	52.45	30.70	70.05	95.48	92.59	95.48	94.82	90.16	94.82	36.68	40.61	40.85
✗	✓	55.17	27.62	67.51	96.64	91.65	96.64	88.60	72.46	88.60	35.47	37.17	38.53
✓	✓	56.44	28.80	69.93	97.08	92.36	97.08	96.07	88.15	96.07	37.86	37.73	37.69

TABLE IV: Ablation studies on competing datasets (Caltech Datasets)

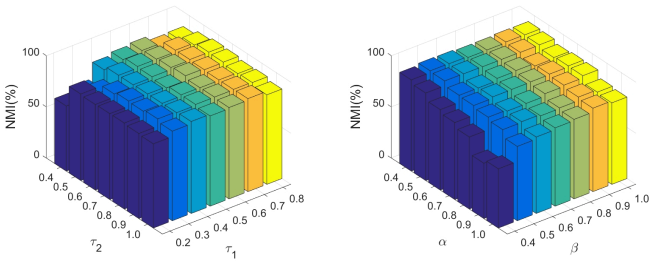
Components		Caltech-2V			Caltech-3V			Caltech-4V			Caltech-5V		
R_{ins}	R_{clu}	**ACC**	**NMI**	**PUR**	**ACC**	**NMI**	**PUR**	**ACC**	**NMI**	**PUR**	**ACC**	**NMI**	**PUR**
✗	✗	60.86	52.85	61.36	60.79	52.92	62.86	83.14	75.37	83.14	85.64	76.71	85.64
✓	✗	62.00	53.91	62.36	66.36	59.85	66.71	89.43	81.80	89.43	85.93	78.94	85.93
✗	✓	60.83	52.95	60.86	66.71	59.36	66.71	86.00	76.55	86.00	85.50	76.39	85.50
✓	✓	66.63	57.62	66.63	68.78	61.27	68.78	89.91	81.66	89.91	86.12	78.94	86.12

metrics: clustering accuracy (ACC), normalized mutual information (NMI), and purity (PUR).

B. Performance

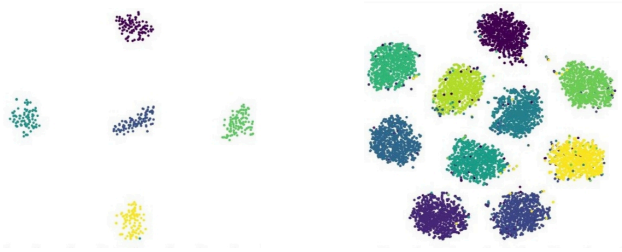
Table I and table II summarizes the clustering performance. We observe that: 1) RCMVC consistently achieves the best results across all metrics. On BBC, it outperforms MFLVC [13] by 9.33% in NMI, mainly due to preserving relational consistency across views through instance-level and cluster-level relational learning. 2) RCMVC shows robust performance across varying views. On the Caltech datasets, it consistently outperforms all comparison methods, demonstrating its effectiveness in capturing cross-view relational information.

Ablation Studies. Ablation studies evaluate the contributions of the instance-level relational contrastive module (R_{ins}) and the cluster-level relational contrastive module (R_{clu}). As shown in Table III and table IV, the full RCMVC model consistently achieves the highest performance by leveraging both components. On the Prokaryotic dataset, the full model improves ACC to 56.44%, a 10.52% gain over the model without both modules (45.92%), demonstrating the contribution of instance-level relational modeling and cluster-level relational modeling.



(a) τ_1 vs τ_2. (b) α vs β.

Fig. 3: Parameter sensitivity analysis on NGs.



NGs Hdigit

Fig. 4: Visualization of the clustering results on representative datasets.

Parameter Sensitivity Analysis. We examine the sensitivity of RCMVC on the NGs dataset with respect to four key hyperparameters in relational contrastive learning: the temperature parameters τ_1 and τ_2, and weighting coefficients α and β. As shown in Fig. 3, NMI changes smoothly with varying hyperparameters, demonstrating that our multi-level relational learning framework is robust to their settings.

Visualization of Clustering Results. We visualize the clustering results using t-SNE on NGs and Hdigit, which reduces the dimension of the extracted feature vectors to 2D. As shown in Fig. 4, distinct cluster structures emerge, further demonstrating the effectiveness of our method.

V. CONCLUSION

In this paper, we propose a novel RCMVC framework for multi-view clustering based on relational contrastive learning. Our approach incorporates relational modeling into contrastive learning to explicitly preserve pairwise similarity consistency and capture both fine-grained instance-level and semantic cluster-level relationships. Extensive experiments on multiple benchmark datasets demonstrate that RCMVC consistently outperforms state-of-the-art methods, highlighting the effectiveness of relational modeling in improving the accuracy of multi-view clustering.

REFERENCES

[1] J. Chen, H. Mao, W. L. Woo, and X. Peng, "Deep multiview clustering by contrasting cluster assignments," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 16752–16761.

[2] J. Xu, Y. Ren, G. Li, L. Pan, C. Zhu, and Z. Xu, "Deep embedded multi-view clustering with collaborative training," *Inf. Sci.*, vol. 573, pp. 279–290, 2021.

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

[4] R. Guo, M. Yang, Y. Lin, X. Peng, and P. Hu, "Robust contrastive multi-view clustering against dual noisy correspondence," in *Adv. Neural Inf. Process. Syst.*, 2024.

[5] X. Wan, J. Liu, H. Yu, Q. Qu, A. Li, X. Liu, K. Liang, Z. Dong, and E. Zhu, "Contrastive continual multiview clustering with filtered structural fusion," *IEEE Trans. Neural Netw. Learn. Syst.*, 2024.

[6] W. Yan, Y. Zhang, C. Tang, W. Zhou, and W. Lin, "Anchor-sharing and clusterwise contrastive network for multiview representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, 2024.

[7] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 8547–8555.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[9] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.

[10] X. Wang and G.-J. Qi, "Contrastive learning with stronger augmentations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5549–5560, 2022.

[11] J. Li, P. Zhou, C. Xiong, and S. C. Hoi, "Prototypical contrastive learning of unsupervised representations," *arXiv preprint arXiv:2005.04966*, 2020.

[12] B. Diallo, J. Hu, T. Li, G. A. Khan, X. Liang, and Y. Zhao, "Deep embedding clustering based on contractive autoencoder," *Neurocomputing*, vol. 433, pp. 96–107, 2021.

[13] J. Xu, H. Tang, Y. Ren, L. Peng, X. Zhu, and L. He, "Multi-level feature learning for contrastive multi-view clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16051–16060.

[14] W. Yan, Y. Zhang, C. Lv, C. Tang, G. Yue, L. Liao, and W. Lin, "Gcfagg: Global and cross-view feature aggregation for multi-view clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19863–19872.

[15] J. Song, H. Zhang, X. Li, L. Gao, M. Wang, and R. Hong, "Self-supervised video hashing with hierarchical binary auto-encoder," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3210–3221, 2018.

[16] S. F. Hussain, C. Grimal, and G. Bisson, "An improved co-similarity measure for document clustering," in *Int. Conf. Mach. Learn. Appl.*, Dec. 2010.

[17] Z. Yu, T. Li, J. Zhao, and J. Luo, "PGAAS: a prokaryotic genome assembly assistant system," *Bioinformatics*, vol. 18, no. 5, pp. 661–665, 2002.

[18] D. Greene and P. Cunningham, "Practical solutions to the problem of diagonal dominance in kernel document clustering," in *Proc. Int. Conf. Mach. Learn.*, 2006, pp. 377–384.

[19] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2004, p. 178.

[20] N. Ali, B. Zafar, F. Riaz, *et al.*, "A hybrid geometric spatial image representation for scene classification," *PLoS One*, vol. 13, no. 9, p. e0203339, 2018.

[21] M.-S. Chen, J.-Q. Lin, X.-L. Li, *et al.*, "Representation learning in multi-view clustering: A literature review," *Data Sci. Eng.*, vol. 7, no. 3, pp. 225–241, 2022.

[22] G. Ke, Z. Hong, Z. Zeng, Z. Liu, Y. Sun, and Y. Xie, "CONAN: contrastive fusion networks for multi-view clustering," in *IEEE Int. Conf. Big Data*, 2021, pp. 653–660.

[23] R. Zhou and Y.-D. Shen, "End-to-end adversarial-attention network for multi-modal clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14619–14628.

[24] D. J. Trosten, S. Lokse, R. Jenssen, and M. Kampffmeyer, "Reconsidering representation alignment for multi-view clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1255–1265.

[25] X. Chen, Y. Ren, J. Xu, F. Lin, X. Pu, and Y. Yang, "Bridging gaps: Federated multi-view clustering in heterogeneous hybrid views," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 37020–37049, 2024.

[26] C. Luo, J. Xu, Y. Ren, J. Ma, and X. Zhu, "Simple contrastive multi-view clustering with data-level fusion," in *Proc. Int. Joint Conf. Artif. Intell.*, 2024, pp. 4697–4705.