Mining Data Patterns In Chinese-English Translation Via Multi-granularity Contrastive Learning

Baoying Yang

School of Foreign Languages, Zhengzhou University of Science and Technology, Zhengzhou, 450064, China

Corresponding author. E-mail: byyang2024@163.com

Received: Jan 18, 2025; Accepted: Mar 13, 2025

Multi-view clustering-based multilingual data pattern mining has received significant attention in recent years due to its ability to fully leverage the complementary and consistent information from multiple languages. Although existing methods achieve encouraging performance, they often jointly optimize representation learning and pattern mining within a single feature space, which may degrade the effectiveness of multilingual data pattern mining. To address this issue, this paper proposes a multi-granularity contrastive learning-based deep multilingual data pattern mining method (MCL), which consists of three view-invariant learning modules: structure learning, semantics learning, and partitioning learning. MCL integrates these three levels of view-invariant learning into an end-to-end framework, comprehensively exploiting the consistency and complementarity of multi-view data, thereby significantly improving the accuracy and robustness of multilingual data pattern mining. Finally, through extensive experiments on five datasets, MCL shows to establish a new benchmark for ACC, NMI, and PUR, proving its superiority and effectiveness.

Keywords: Multi-granularity Contrastive Learning; tri-invariant alignment; multilingual data mining

http://dx.doi.org/10.6180/jase.202511_28(11).0019

1. Introduction

With the acceleration of globalization and the widespread adoption of digital technologies, multilingual data has experienced explosive growth on the internet, social media, corporate documents, and cross-cultural communications [1, 2]. This multilingual nature reflects the diversity of human society while also presenting new challenges for data-driven research and applications. Traditional single-language data mining methods are often limited to the grammatical structures, semantic expressions, and cultural contexts of specific languages, making them difficult to directly adapt to multilingual scenarios. For instance, English-centric models may fail when processing morphologically rich languages (e.g., Arabic) or character-based languages (e.g., Chinese) due to incompatible feature representations. Against this backdrop, multilingual data pattern mining has emerged, aiming to extract universal or language-specific latent patterns from heterogeneous language data through cross-lingual alignment, transfer learning, and semantic generalization techniques. These efforts support tasks such as cross-lingual information retrieval, machine translation, sentiment analysis, and knowledge discovery [3].

In recent years, deep learning-based multi-view clustering methods have made significant progress in mining multilingual data patterns, via considering each language as a view [4]. Initially, deep clustering methods combined autoencoders with clustering algorithms to alternately or simultaneously perform feature representation learning and clustering, achieving better results than heuristic methods. For examples, Xu et al. utilized autoencoders to learn the embedding representations of multiple views and then sequentially trained them to collaboratively extract consistent

complementary information [5, 6]. Xiao et al. applied graph convolutional networks to each view to capture complementarities and learn discriminative representations, then adaptively combined attribute and structural information across different views to capture consistent information [7, 8]. Additionally, contrastive learning, due to its unsupervised nature, has been explored for learning discriminative feature representations that differentiate clusters without labels, effectively integrating multi-view information and achieving superior performance [8–11].

Despite the significant advancements made by existing methods in deep contrastive multi-view clustering, they still faces two critical challenges exacerbated by the heterogeneous nature of cross-lingual environments: (1) Objective conflicts in single latent space: they often enforce multiple constraint objectives (e.g., cross-view alignment, discriminative feature learning) within a single latent space. However, these objectives may inherently conflict due to their divergent optimization goals. For instance, features optimized for one task (e.g., view-specific reconstruction) might act as noise for another task (e.g., cross-view consistency), leading to compromised discriminative power and suboptimal clustering. Existing approaches that naively combine losses in a shared representation layer fail to resolve this paradoxical interference, resulting in entangled and task-conflicting features. (2) Current methods typically treat cross-view pairs of the same sample as positives and all other pairs (even semantically similar samples within the same view) as negatives. This rigid binary contrastive strategy ignores intrinsic intra-view similarities, forcing dissimilarity between samples that should belong to the same cluster. Consequently, the learned representations push semantically related samples apart within the same view, violating cluster cohesion and undermining cross-view consistency. This dual failure—intra-view fragmentation and inter-view misalignment—directly harms the model's ability to recover unified cluster structures across views.

To address these challenges, a multi-granularity contrastive learning based deep multi-view clustering (MCL) is proposed for mining multilingual data patterns, which consists of the view-invariant structure learning, the view-invariant semantics learning, and the view-invariant partitioning learning. Specifically, MCL conducts view-invariant structure learning to align low-level features of multi-view data within the encoder-decoder architecture for learning inherent information of each samples. Then, MCL conducts the view-invariant semantic learning by introducing global structural relationships as soft weights for negative samples to enhance intra-class compactness. Meanwhile, MCL conducts the view-invariant partitioning learning via maximizing the mutual information between cluster distributions between views to ensure the partitioning consistency. The three levels of granularity in contrastive learning collaborate in a non-fusion manner within an end-to-end framework, exploring the consistency and complementarity information in multi-view data more comprehensively. Finally, through extensive experiments on five datasets, MCL shows to establish a new benchmark for ACC, NMI, and PUR, proving its superiority and effectiveness.

The contributions of this paper are threefold:

- A deep multi-granularity multi-view contrastive clustering is proposed via conducting the view-invariant structure learning, the view-invariant semantic learning, and the view-invariant partitioning learning within different spaces, which effectively mines multilingual data patterns.

- An adaptive structure contrastive loss is designed to learn semantics divergence between multi-view data, which enhances intra-cluster compactness and inter-cluster separability.

- Numerous results on five datasets from different fields demonstrate MCL sets a new cutting-edge baseline in the multilingual data mining task.

The structure of the paper is as follows: Section 2 provides a detailed description of MCL; Section 3 evaluates the performance of MCL on several standard datasets, with a comprehensive comparison to existing techniques; Section 4 summarizes the key contributions of this work and outlines directions for future research.

2. Methodology

Given a multilingual dataset $\{x_i^1, x_i^2, \ldots, x_i^V\}_{i=1}^n$, where V is the number of languages, n is the number of samples, and $x_i^v \in R^{D_v}$ represents the i-th sample from the v-th language type, D_v denotes the data dimension of the v-th language type. Multilingual pattern mining aims to group data samples into K meaningful categories by leveraging the similarity between cooperative samples and the difference between categories without requiring manual annotations. To achieve this goal, a multi-granularity contrastive learning-based deep multilingual data pattern mining method (MCL) is proposed, which consists of the view-invariant structure learning, the view-invariant semantics learning, and the view-invariant partitioning learning, as shown in Fig. 1.

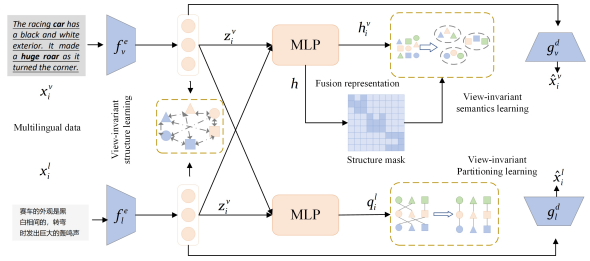

Fig. 1. The illustration of MCL, containing the view-invariant structure learning, the view-invariant semantics learning, and the view-invariant partitioning learning.

2.1. View-invariant structure learning

View-invariant structure learning aims to align low-level features of multi-view data within the encoder-decoder architecture via treating multiple views of samples as the positive pairs and multiple views of other samples as the negative pairs, to learn inherent information of each sample.

Specifically, given the multi-view dataset with n samples and V views $\{X^v\}_{v=1}^V$, the V pairs of the encoder and decoder are used to map the multi-view data into the latent space to extract low-level representations:

$$z_i^v = f_v^e(x_i^v), \quad \hat{x}_i^v = g_v^d(z_i^v) \tag{1}$$

where $f_v^e(\cdot)$ and $g_v^d(\cdot)$ denote the v-th view encoder and decoder, respectively. z_i^v and $\hat{x}_i^v$ denote low-level representations and reconstruction data of the corresponding sample, respectively. Then, the reconstruction loss is used to optimize the overall network:

$$L_{\text{rec}} = \sum_{v=1}^{V} L_{\text{rec}}^v = \sum_{i=1}^{n} \|x_i^v - g_v^d(f_v^e(x_i^v))\|_2^2 \tag{2}$$

where $\|\cdot\|_2$ denotes the L_2 norm. Then, a view-invariant structure contrastive loss is designed to maximize the structure consistency between low-level representations:

$$L_s = \frac{1}{2}\sum_{v=1}^{V}\sum_{l\neq v}^{V} \mathcal{H}(\mathbf{T_s}^{(v,l)}, \cos(z^v, z^l)) + \mathcal{H}(\mathbf{T_s}^{(v,l)}, \cos(z^v, z^v)) \tag{3}$$

where $\mathcal{H}$ denotes the the cross entropy. T_s is the structure mask matric that indicates the given pair is positive or negative between the v-th view and the l-th view. $\cos(\cdot)$ denotes the cosine similarity whose definition is shown as follows:

$$\cos(z_i^l, z_j^v) = \frac{\langle z_i^l, z_j^v \rangle}{\|z_i^l\| \|z_j^v\|} \tag{4}$$

2.2. View-invariant semantics learning

After learning the inherent structural information of each sample, a view-invariant semantic learning approach is designed by introducing global structural relationships as soft weights for negative samples to enhance intra-class compactness.

Specifically, to avoid conflicts between the objectives of view-invariant structural learning and view-invariant semantic learning, a view-shared multi-layer perceptron (MLP) network is employed to transform low-level structure representations into high-level semantic representations.

$$h_i^v = MLP(z_i^v, W_s) \tag{5}$$

where W_s denotes the network parameters. Then, high-level semantic representations h^v from all views are concatenated to obtain high-level semantic fusion representations $\mathbf{h}$:

$$\mathbf{h} = Fusion(h^1, h^2, \ldots, h^V) \tag{6}$$

where $h^v = [h_1^v, h_2^v, \ldots, h_n^v] \in \mathbb{R}^{n\times d_v}, h_i^v \in \mathbb{R}^{1\times d_v}, \mathbf{h} \in \mathbb{R}^{n\times d}, d = d_v \times V$. Therefore, the structure relationship between samples is computed as:

$$\mathbf{s} = \frac{\mathbf{h}\mathbf{h}^\top}{K} \tag{7}$$

where $\mathbf{s}$ is the similarity matrix and each element s_{ij} represents the similarity between sample i and sample j. And K is the scaling factor. The view-invariant semantics contrastive learning is defined as follows:

$$L_g = -\frac{1}{2N}\sum_{i=1}^{n}\sum_{v=1}^{V}\sum_{l=1}^{V} \log \frac{e^{\cos(h_i^v, h_i^l)/\tau}}{\sum_{j=1}^{n} e^{(1-s_{ij})\cos(h_i^v, h_i^l)/\tau} - e^{1/\tau}} \tag{8}$$

This loss function leverages contrastive learning principles to enhance the model's performance by incorporating sample similarity as dynamic weights. By explicitly encouraging high similarity between positive pairs, it ensures tighter intra-class compactness in the feature space. At the same time, the adaptive weighting of negative samples, guided by their similarity scores, effectively reduces the influence of noisy negatives while emphasizing hard negatives. Additionally, the inclusion of a temperature parameter τ allows for flexible control over the sensitivity of the contrastive learning process, making it adaptable to various data distributions and tasks. Overall, this function improves the model's ability to learn robust and discriminative representations while preserving the global structural relationships among samples.

2.3. View-invariant partitioning learning

Different views are different descriptions of the same sample and should adhere to a consistent cluster assignment.

To this end, a view-invariant partitioning learning is devised via maximizing the mutual information between cluster distributions between views to ensure the partitioning consistency.

Specifically, a view-shared clustering head is devised with the help of the MLP network to extract soft cluster assignments across views:

$$q^v = MLP(h^v, W_c) \tag{9}$$

where W_c denotes the network parameters of the clustering head. Then, the mutual information $I(q^v, q^l)$ between inter-view cluster partitioning is maximizing via:

$$\begin{aligned} \max(q^{(v)}, q^{(l)}) = & \iint p(q^{(l)}|q^{(v)})p(q^{(v)}) \\ & \log \frac{p(q^{(l)}|q^{(v)})}{p(q^{(l)})} \, \mathrm{d}q^{(v)} \, \mathrm{d}q^{(l)} \\ & = KL(p(q^{(l)}|q^{(v)})p(q^{(v)}) \\ & \parallel p(q^{(l)})p(q^{(v)})) \end{aligned} \tag{10}$$

where $KL(\cdot \parallel \cdot)$ denotes the Kullback-Leibler (KL) divergence. While KL divergence is a widely used metric for measuring the dissimilarity between probability distributions, it is inherently unbounded and can lead to instability during optimization. To address this issue, the Jensen-Shannon (JS) divergence is used as a substitute for KL divergence. JS divergence is a symmetric and bounded measure, offering better numerical stability and interpretability. The reformulated optimization objective is given by:

$$\begin{aligned} \max I(q^{(v)}, q^{(l)}) = & JS(p(q^{(l)} \mid q^{(v)})p(q^{(v)}) \\ & \parallel p(q^{(l)})p(q^{(v)})) \end{aligned} \tag{11}$$

By leveraging JS divergence, the model ensures robust optimization of the mutual information, facilitating effective alignment of inter-view cluster assignments. This framework captures shared semantics across views, preserves structural consistency, and mitigates the impact of noisy or redundant features in individual views, leading to a more unified and interpretable representation. Based on the variational estimation of JS divergence,

$$\begin{aligned} \max I(q^{(v)}, q^{(l)}) = & \mathbb{E}_{(q^{(v)}, q^{(l)}) \sim p\left(q^{(l)}|q^{(v)}\right)p(q^{(v)})} \\ & \left[\log \rho \left(T\left(q^{(v)}, q^{(l)}\right)\right)\right] \\ & + \mathbb{E}_{(q^{(v)}, q^{(l)}) \sim p\left(q^{(l)}\right)p(q^{(v)})} \\ & \left[\log\left(1 - \rho\left(T\left(q^{(v)}, q^{(l)}\right)\right)\right)\right] \end{aligned} \tag{12}$$

To practically achieve mutual information maximization and enhance consistent learning across multi-view data, negative sample estimation is applied. In this approach, positive and negative pairs of data are constructed based on their respective representations. The function $\rho(T())$ serves as a discriminator, distinguishing between positive and negative sample pairs. This process enables the model to maximize the correlation between views, ensuring that shared semantics are effectively captured while promoting robust feature alignment across different views. Meanwhile, to avoid trivial solutions, define a cluster entropy maximization loss:

$$L_c = -\sum_{v=1}^{V}\sum_{k=1}^{K}\left(\frac{1}{n}\sum_{i=1}^{n} q_{ik}^v\right)\log\left(\frac{1}{n}\sum_{i=1}^{n} q_{ik}^v\right) \tag{13}$$

Finally, the semantic label of the i-th sample is calculated via:

$$y_i = \arg\max_k \left(\frac{1}{V}\sum_{v=1}^{V} q_{ik}^v\right) \tag{14}$$

2.4. The overall objective function

MCL utilizes the loss L to train the overall architecture with stochastic gradient descent optimization algorithm:

$$L = L_{rec} + \alpha L_s + \beta L_g - \gamma max I(q^{(v)}, q^{(l)}) + L_c \tag{15}$$

where α, β and γ are hyperparameters that are used to balance three view-invariant learning. The detailed training process is shown in the table 1. The optimization of the objective function in Eq. 16 is carried out using stochastic gradient descent (SGD). At each training iteration, a mini-batch of samples is randomly selected from the multi-view dataset. For this mini-batch, the reconstruction loss L_{rec} the structure-preserving loss L_s, the global relationship loss L_g, and the mutual information term $I(q^{(v)}, q^{(l)})$ are computed. These components are combined using the hyperparameters α, β, and γ to compute the total loss L as defined in Eq. 16. The model parameters θ are then updated using the gradient of the loss function with respect to the parameters, following the update rule:

$$\theta \leftarrow \theta - \eta \frac{\partial L}{\partial \theta} \tag{16}$$

where η denotes the learning rate. This process is repeated iteratively until the stopping criterion is met, such as the predefined number of iterations or the convergence of the loss function. The use of SGD ensures efficient optimization of the objective function, enabling the model to learn view-invariant features that are effective for clustering. The detailed information is shown in Table 1.

Table 1. Algorithm Flow of MCL

Input	Multi-view dataset $X = \{x_i^1, x_i^2, \ldots, x_i^V\}_{i=1}^n$, number of training iterations N, hyperparameters α, β, γ.
Output	Clustering results $\{y_i^n\}_{i=1}$.
Loop	Repeat the following steps until the number of iterations reaches N:
Step 1	Obtain low-level structure representations z^v using multi-view autoencoders.
Step 2	Obtain high-level semantic features h^v through the MLP network.
Step 3	Compute the fused representation **h**.
Step 4	Obtain the structure relationship matrix **s**.
Step 5	Compute the reconstruction loss L_{rec} using Eq. (2).
Step 6	Compute the view-invariant structure learning loss L_s using Eq. (3).
Step 7	Compute the view-invariant semantics learning loss L_g using Eq. (8).
Step 8	Compute the view-invariant partitioning learning using Eq. (12).
Step 9	Compute the combined loss L using Eq. (14).
Step 11	Optimize the MCL using the stochastic gradient descent strategy.

Table 2. The detailed statistics of five datasets.

Datasets	BDGP	CCV	Caltech-5V	CLS	MARC
Class	5	2000	7	2	5
Sample	20	6773	1440	4000	2000
View	2	3	5	4	6

3. Results and discussion

3.1. Setup

Dataset and Metric: Five datasets are utilized in the experiments [10, 11]. The detailed dataset statistics are shown in Table 2. BDGP dataset is collected from the berkeley drosophila genome project and contains 2500 samples forming two views of the dataset. Columbia Consumer Video (CCV) dataset is an internet video dataset that encompasses a wide range of real-world daily life scenes such as family gatherings, travel, street scenes, and sports activities, with 6773 samples across 20 categories. Caltech-5V dataset is a multi-view RGB image dataset, with 1440 samples across 7 categories. CLS dataset is a cross-lingual sentiment classification dataset with four language views, containing 4000 samples across 2 categories. MARC is an Amazon product review dataset with six language views, containing 2000 samples across 5 categories. ACC, NMI, and PUR are used to validate the performance between methods. ACC measures the degree of alignment between the model's predictions and the true labels, representing the proportion of correctly classified samples out of the total samples. NMI measures the information shared between the clustering results and the true labels, with higher values indicating stronger consistency between the clustering results and the true categories. PUR evaluates the purity of the clustering by calculating the proportion of the most frequent true category in each cluster, reflecting the effectiveness of the clustering.

Implementation Details: All the experiments are implemented by Python on the RTX 3090 GPU . In the experiments, all datasets are reshaped into vectors, and autoencoders for all views are implemented using fully connected networks with similar architectures. Each view's encoder follows a four-layer fully connected structure: Input_dim $\rightarrow$ Fc_500 $\rightarrow$ Fc_500 $\rightarrow$ Fc_2000 $\rightarrow$ Fc_512, while the decoder mirrors the encoder's architecture symmetrically. The latent representations for all views are then transformed to a dimensionality of 512. Both the encoder and decoder employ the ReLU activation function. Additionally, for each view, a single-layer MLP extracts high-level features, and another single-layer MLP with a Softmax layer constructs the clustering predictor, generating pseudo-labels for the multiple views. The output dimension of the clustering predictor corresponds to the number of clusters. To enhance robustness, normalization is applied prior to the Softmax output. The output dimension of the feature MLP is set to 128. The parameters α, β and γ are set to 0.1, 1, and 0.1 on all datasets, respectively. The learn rate is set in [0.0001, 0.001]. The number of the batch size and epoch is set 512 and 500, respectively.

3.2. Comparison with baselines

Comparison baselines: Ten methods are compared on three datasets about ACC, NMI, and PUR, to demonstrate the performance of MCL, containing GCFAgg [12], AFMVC [13], DCP [14], AFI [15], SSMVC [16], MDCN [17], DAMVSI [18], MJL [2], ILSK [19], and PTMTC [20].

Comparison results: In the experiment, MCL is comprehensively compared with existing methods. The experiment results in the Table 3 and Fig. 2 show that MCL achieves significant advantages across multiple evaluation metrics. The reason are threefold: (1) MCL aligns low-level features across different views without early fusion. This preserves the local structure and fine-grained details of each view, ensuring that the fundamental relationships within each sample remain intact. By mapping features

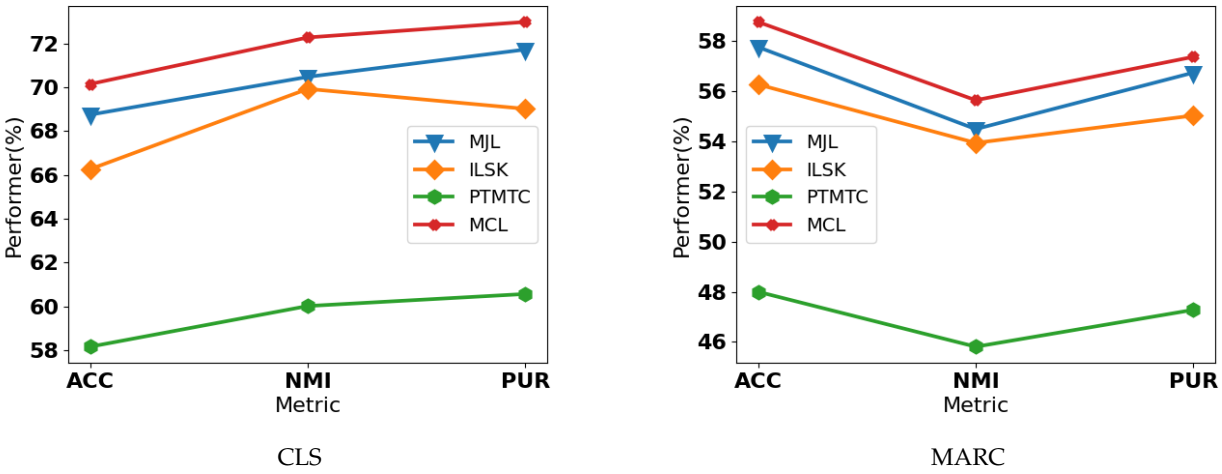

Fig. 2. Comparison results on the multilingual datasets.

Table 3. Comparison results on the BDGP, CCV, and Caltech-5V datastets.

	BDGP			CCV			Caltech-5V		
Method	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
GCFAgg	0.9859	0.9584	0.9339	0.1645	0.2231	0.1004	0.5859	0.5314	0.5639
AFMVC	0.9405	0.9412	0.9456	0.1676	0.2004	0.1445	0.5450	0.5212	0.5458
DCP	0.9491	0.9213	0.9577	0.2212	0.2525	0.2877	0.6491	0.5513	0.6530
AFI	0.9746	0.9457	0.9745	0.2275	0.2664	0.2705	0.6246	0.5012	0.5659
SSMVC	0.9796	0.9458	0.9777	0.2338	0.2442	0.2722	0.5596	0.5300	0.5514
MDCN	0.9378	0.9123	0.9235	0.2258	0.2413	0.2522	0.5090	0.4723	0.5893
DAMV-SI	0.9858	0.9512	0.9858	0.2558	0.2412	0.2978	0.6483	0.5705	0.6498
MCL	**0.9860**	**0.9517**	**0.9860**	**0.2643**	**0.2765**	**0.3236**	**0.6621**	**0.5775**	**0.6621**

to a shared space while maintaining the individuality of each view, MCL lays a solid foundation for multi-view clustering. (2) MCL refines inter-sample relationships using a contrastive learning framework with soft weights for negative samples. This enhances intra-class compactness and inter-class separability, leading to more accurate clustering. By focusing on global structural relationships, this phase strengthens class representations and ensures that similar samples across views are grouped together effectively. (3) MCL maximizes mutual information between cluster distributions across views, reinforcing partition consistency. This ensures that clustering structures remain coherent across different views, preserving their complementary information. The absence of explicit view fusion makes MCL adaptable to diverse and heterogeneous data, improving its robustness and generalization.

3.3. Parameter analysis

To evaluate the impact of the hyperparameters α, β, and γ, which control the balance between different loss components in MCL, we conducted a detailed parameter analysis on a single dataset in terms of ACC and NMI. The values of α, β, and γ were varied within the range $\{10, 1, 0.1, 0.01\}$. The results show that MCL achieves optimal performance when these hyperparameters are set within the range $\{0.1, 1\}$, demonstrating their critical role in balancing multi-view complementary learning and the robustness of the model under reasonable settings. Specifically, these hyperparameters ensure an effective trade-off among the objectives of concentration, consistency, and comprehensiveness, thereby improving the quality of representation learning and clustering performance. However, when any of the hyperparameters is set too high or too low (e.g., 10 or 0.01), the model's performance decreases significantly, indicating that overly strong or weak contributions from a single loss component hinder the effective integration of multi-view information. Overall, these findings highlight the importance of carefully tuning α, β, and γ, while further validating the robustness and effectiveness of the MCL framework in handling multi-view data.

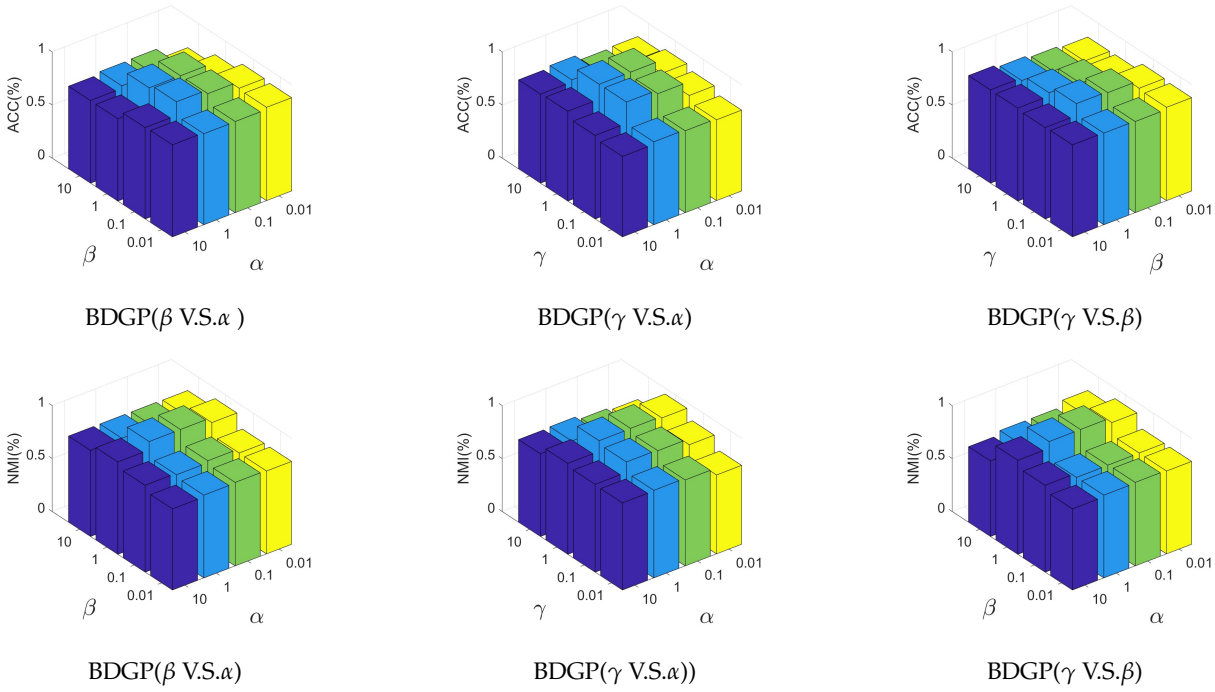

Fig. 3. The hyperparameter analysis of α, β and γ on the BDGP dataset.

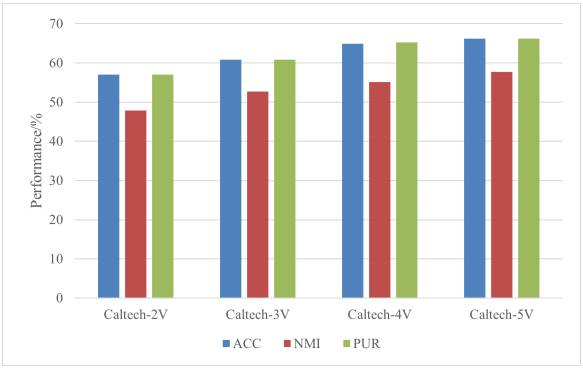

Fig. 4. Multi-view fusion clustering performance on the Caltech-5V with multiple views.

3.4. View analysis

Based on the results in Fig. 4, it is evident that the performance of clustering improves consistently as the number of views increases in the Caltech-5V dataset. Specifically, the metrics ACC, NMI, and PUR all show a clear upward trend from Caltech-2V to Caltech-5V. This highlights the effectiveness of multi-view data fusion in improving clustering performance. When only two views are utilized (Caltech-2V), the ACC, NMI, and PUR values are relatively low, at 57, 47.88, and 57, respectively. As the number of views increases to three (Caltech-3V), the metrics improve significantly, with ACC reaching 60.86, NMI increasing to 52.66, and PUR also rising to 60.86. This indicates that adding an additional view provides complementary information that enhances clustering consistency and accuracy. With four views (Caltech-4V), the performance metrics further improve to 64.86 for ACC, 55.16 for NMI, and 65.21 for PUR. This continued improvement suggests that incorporating more views enables a more comprehensive understanding of the data, capturing both shared and unique information from different perspectives. Finally, when all five views are integrated (Caltech-5V), the metrics reach their highest values, with ACC at 66.21, NMI at 57.75, and PUR at 66.21. This demonstrates that the fusion of multiple views maximizes the utilization of diverse information, leading to more robust and consistent clustering outcomes. In summary, the progressive improvement across all metrics with an increasing number of views validates the benefits of multi-view data fusion. By integrating complementary information from multiple views, the clustering performance achieves significant enhancements, demonstrating the importance of leveraging the full potential of multi-view data for accurate and reliable clustering.

4. Conclusion

This paper proposed the Multi-granularity Contrastive Learning (MCL) framework for mining multimedia image patterns. The proposed MCL integrates three levels of view-invariant learning: structure, semantics, and par-

titioning, within an end-to-end framework. By addressing key challenges such as conflicting constraints in a single latent space and inconsistent sample representations, MCL demonstrated robust and consistent clustering performance across diverse datasets. Extensive experiments validated its superiority in terms of ACC, NMI, and PUR, setting new benchmarks in multi-view clustering tasks. The parameter and view analysis further confirmed the framework's robustness and the effectiveness of multi-view fusion. These findings underscore the potential of MCL as a powerful tool for multi-view data mining, paving the way for more accurate and comprehensive insights in multimedia analysis. We will focus on adapting MCL to dynamic multi-view clustering to better handle evolving data streams, integrating weakly-supervised learning to enhance clustering accuracy with limited labeled data, and developing cross-modal fusion capabilities to leverage diverse data types. Additionally, we plan to improve scalability through graph-based methods and address the challenges of noisy or missing views to strengthen the robustness of the framework.

References

[1] X. Yan, H. Huang, Y. Jin, L. Chen, Z. Liang, and Z. Hao, (2023) *"Neural architecture search via multi-hashing embedding and graph tensor networks for multilingual text classification"* **IEEE Transactions on Emerging Topics in Computational Intelligence** *8*(1): 350–363. DOI: 10.1109/TETCI.2023.3301774.

[2] A. Ekbal et al., (2024) *"Atmosphere kamaal ka tha (was wonderful): A multilingual joint learning framework for aspect category detection and sentiment classification"* **IEEE Transactions on Computational Social Systems**: DOI: 10.1109/TCSS.2024.3374450.

[3] J. Gao, P. Li, A. A. Laghari, G. Srivastava, T. R. Gadekallu, S. Abbas, and J. Zhang, (2024) *"Incomplete multiview clustering via semidiscrete optimal transport for multimedia data mining in IoT"* **ACM Transactions on Multimedia Computing, Communications and Applications** ***20***(6): 1–20. DOI: 10.1145/3625548.

[4] J. Xu, H. Tang, Y. Ren, L. Peng, X. Zhu, and L. He. "Multi-level feature learning for contrastive multi-view clustering". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, 16051–16060.

[5] S. Xiao, S. Du, Z. Chen, Y. Zhang, and S. Wang, (2023) *"Dual fusion-propagation graph neural network for multi-view clustering"* **IEEE Transactions on Multimedia** ***25***: 9203–9215. DOI: DOI:10.1109/TMM.2023.3248173.

[6] L. Fu, S. Huang, L. Zhang, J. Yang, Z. Zheng, C. Zhang, and C. Chen, (2024) *"Subspace-contrastive multi-view clustering"* **ACM Transactions on Knowledge Discovery from Data** ***18***(9): 1–35. DOI: 10.1145/367483.

[7] Y. Sun, Y. Qin, Y. Li, D. Peng, X. Peng, and P. Hu, (2024) *"Robust multi-view clustering with noisy correspondence"* **IEEE Transactions on Knowledge and Data Engineering**: DOI: 10.1109/TKDE.2024.3423307.

[8] J. Gao, M. Liu, P. Li, A. A. Laghari, A. R. Javed, N. Victor, and T. R. Gadekallu, (2023) *"Deep incomplete multi-view clustering via information bottleneck for pattern mining of data in extreme-environment IoT"* **IEEE Internet of Things Journal**: DOI: 10.1109/JIOT.2023.3325272.

[9] S. Shi, F. Nie, R. Wang, and X. Li, (2021) *"Multi-view clustering via nonnegative and orthogonal graph reconstruction"* **IEEE transactions on neural networks and learning systems** ***34***(1): 201–214. DOI: 10.1109/TNNLS.2021.3093297.

[10] P. Zhang, S. Wang, L. Li, C. Zhang, X. Liu, E. Zhu, Z. Liu, L. Zhou, and L. Luo. "Let the data choose: Flexible and diverse anchor graph fusion for scalable multi-view clustering". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. ***37***. 9. 2023, 11262–11269. DOI: 10.1609/aaai.v37i9.26333.

[11] Y. Ren, J. Pu, C. Cui, Y. Zheng, X. Chen, X. Pu, and L. He. "Dynamic weighted graph fusion for deep multi-view clustering". In: *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*. 2024, 4842–4850.

[12] 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: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 19863–19872.

[13] R. Zhang, S. Hang, Z. Sun, F. Nie, R. Wang, and X. Li, (2025) *"Anchor-based fast spectral ensemble clustering"* **Information Fusion** ***113***: 102587. DOI: 10.1016/j.inffus.2024.102587.

[14] Y. Lin, Y. Gou, X. Liu, J. Bai, J. Lv, and X. Peng, (2022) *"Dual contrastive prediction for incomplete multi-view representation learning"* **IEEE Transactions on Pattern Analysis and Machine Intelligence** ***45***(4): 4447–4461. DOI: 10.1109/TPAMI.2022.3197238.

[15] J. Pu, C. Cui, X. Chen, Y. Ren, X. Pu, Z. Hao, S. Y. Philip, and L. He. "Adaptive Feature Imputation with Latent Graph for Deep Incomplete Multi-View Clustering". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. ***38***. 13. 2024, 14633–14641. DOI: 10.1609/aaai.v38i13.29380.

[16] S. Wang, X. Liu, S. Liu, W. Tu, and E. Zhu, (2024) *"Scalable and structural multi-view graph clustering with adaptive anchor fusion"* **IEEE Transactions on Image Processing**: DOI: 10.1109/TIP.2024.3444320.

[17] P. Li, A. A. Laghari, M. Rashid, J. Gao, T. R. Gadekallu, A. R. Javed, and S. Yin, (2022) *"A deep multimodal adversarial cycle-consistent network for smart enterprise system"* **IEEE Transactions on Industrial Informatics *19***(1): 693–702. DOI: 10.1109/TII.2022.3197201.

[18] J. Gao, M. Liu, P. Li, J. Zhang, and Z. Chen, (2024) *"Deep Multiview Adaptive Clustering With Semantic Invariance"* **IEEE Transactions on Neural Networks and Learning Systems *35***(9): 12965–12978. DOI: 10.1109/TNNLS.2023.3265699.

[19] H. Kanayama, Y. Zhao, R. Iwamoto, and T. Ohko. "Incorporating syntax and lexical knowledge to multilingual sentiment classification on large language models". In: *Findings of the Association for Computational Linguistics ACL 2024*. 2024, 4810–4817.

[20] C. Wang and M. Banko. "Practical transformer-based multilingual text classification". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*. 2021, 121–129.