Multi-granularity Self-learning Guided Deep Multi-view Clustering

Zhikui Chen*, Yuzhe Li
School of Software Technology, Dalian University of Technology, Dalian, China
zkchen@dlut.edu.cn, 1594873855@mail.dlut.edu.cn

***Abstract*—Although deep multi-view clustering achieves encouraging performance, there still remain two issues. (1) Most methods commonly utilize inter-view invariance to aggregate information across views, which results in the loss of useful complementary information in each view. (2) Most methods do not fully model inherent self-learning signals, resulting in suboptimal clustering results. To this end, this paper proposes a Multi-granularity Self-learning Guided Multi-View Clustering (MSG-MVC). Specifically, MSG-MVC devises a sample-guided information compression module to capture discriminant view-specific representations via self-learning compression with sample manifold preservation. Then, it designs representation-guided complementary integration to learn semantics-robust fusion representations via self-learning transformation with the conditional entropy minimization. Meanwhile, it introduces a cluster-guided consistent alignment module to capture a consensus cluster assignment from representations via self-learning partition with the multi-level semantic invariance. Experiments on eight datasets show that MSG-MVC achieves cutting-edge performance. The code is available at https://github.com/Yuzhe-Li123/MSG-MVC.**

***Index Terms*—multi-view clustering, complementary information fusion, multi-granularity self-learning**

I. Introduction

Multi-view data, which characterizes samples from multiple perspectives, is ubiquitous in various domains, such as cross-modal retrieval [1], pattern recognition [2], and autonomous driving [3]. Deep multi-view clustering (DMVC) can effectively mine patterns of data without labels via exploiting complementary and consensus information embedded in different views, which has become an active area of research [4].

Existing DMVC methods can be divided into two branches: two-stage and one-stage approaches. The former aligns view-specific representations through inter-view invariance to aggregate information of views, and then executes a vanilla clustering strategy on fusion representations. For example, MGBCC partitions data into different granular balls in each view to explore clustering structures via maximizing consistency between similar granular-balls and maximizing discrepancy between dissimilar granular balls [5]. The latter integrates multi-view representation learning and clustering pattern mining into an end-to-end deep network where pseudo-labels of data are used as self-learning signals for guiding data structure exploration. For example, DMAC-SI proposes a dual assignment invariance clustering method that maximizes semantic consistency between view-specific representations, and maximizes partition consistencies between fusion and augmented fusion representations [6].

Although existing DMVC methods have achieved promising results, two challenges remain. (1) Most methods commonly force view-specific representations to be as similar as possible to bridge the heterogeneous gap between views for aggregating information of views, which leads to the loss of useful complementary information in each view and restricts the diversity of view-specific representations. (2) Most methods only consider pseudo-label information of data as self-learning signals and ignore supervision signals in other spaces such as the data space and the representation space, which leads to suboptimal clustering results due to insufficient information extraction.

To tackle the challenges, we propose a Multi-granularity Self-learning Guided Deep Multi-View Clustering (MSG-MVC) method within the one-stage paradigm. Specifically, MSG-MVC devises a sample-guided information compression module that maximizes mutual information between view-specific data and representations, to capture intrinsic information in each view by treating data as self-learning signal. Then, it designs a representation-guided complementary integration module that minimizes conditional entropy between fusion and view-specific representations, to fully aggregate complementary information between views by treating fusion representations as self-learning signal. Meanwhile, it introduces a cluster-guided consistent alignment module that maximizes mutual information between cluster assignments of the fused and the view-specific representations at different levels, to capture a consensus assignment by treating pseudo-labels as self-learning signals. The contributions of MSG-MVC are:

- A bottom-up information-oriented clustering paradigm is proposed via performing self-guided learning at the sample, representation, and cluster levels to fully mine inherent information of multi-view data.
- A novel complementarity–consistency collaborative learning strategy is proposed, which learns robust fusion representations via conditional-entropy minimization to preserve complementary information and enforces semantic consistency through instance-level and cluster-level alignment.
- Extensive experiments on eight datasets show that MSG-MVC achieves cutting-edge performance in terms of three clustering metrics in comparison with nine baselines.

This work was supported by the National Key R&D Program of China under Grant 2025YFF0514800.

*Corresponding author: Zhikui Chen.

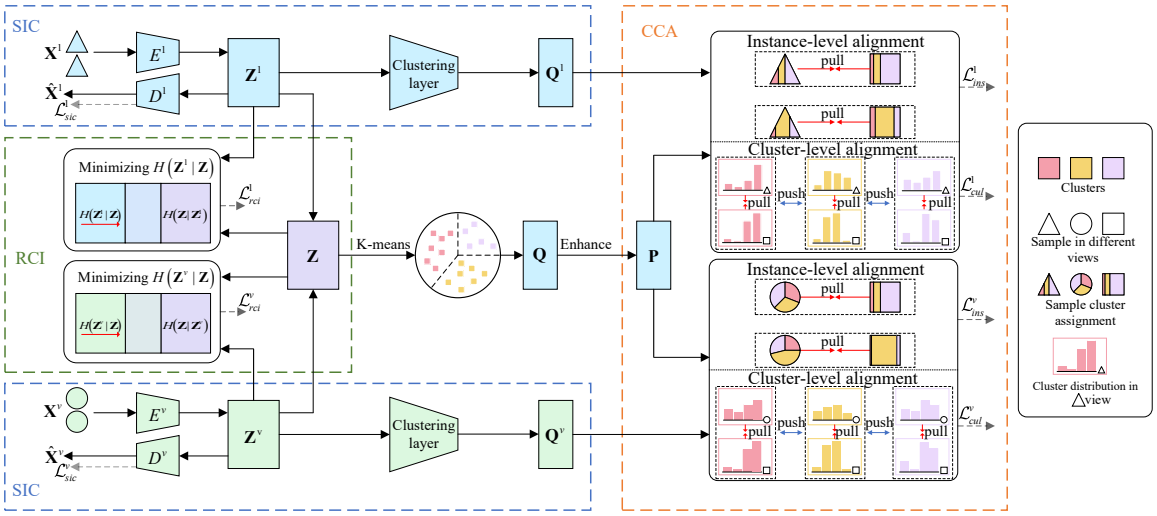


Fig. 1. The illustration of MSG-MVC. SIC utilizes autoencoders to learn the view-specific representation $\mathbf{Z}^v$. RCI minimizes the conditional entropy $H(\mathbf{Z}^v|\mathbf{Z})$ to learn semantics-robust fusion representations. CCA captures a consensus cluster assignment through instance-level and cluster-level semantic invariance.

II. Related Work

Existing DMVC methods can be categorized into two groups. The first group follows a two-stage pipeline: they first aggregate view-specific representations to obtain fusion representations, and then perform clustering on the fusion representations to obtain clustering results. Besides the previously mentioned MGBCC, DGR [7] adaptively fuses multi-view graphs into a consensus graph and trains a residual GCN with reconstruction and orthogonality regularization to obtain clustering-friendly embeddings. DMVC-CE [8] fuses multi-view features via a mixture-of-experts with a gating network and regularizes training to keep experts balanced and diverse. MVC-IIV [9] proposes a two-stage incremental deep multi-view clustering framework that uses embedding and centroid consistency to handle new samples and new views. The above methods decouple representation learning from clustering, thus they cannot utilize self-learning signals derived from clustering results to guide representation learning. To overcome this limitation, one-stage DMVC methods have been developed, which integrates representation learning and clustering end-to-end, with pseudo-labels providing self-training guidance. In addition to DMAC-SI, TMCC [10] learns view-consensus representations using an easy-to-complex adaptive filter and contrastive learning over inter-modal and intra-modal similarities. SWIB [11] fuses multi-view information and learns view weights using both view quality and pseudo-label self-supervision. DFMVC-AKAN [12] employs an attention-enhanced KAN to learn multi-view representations and a distribution-alignment constraint to achieve fair clustering.

III. Method

Given a multi-view dataset with N samples and V views $\{\mathbf{x}_i\}_{i=1}^N = \{\mathbf{x}_i^1 \in \mathbb{R}^{d^1}, ..., \mathbf{x}_i^V \in \mathbb{R}^{d^V}\}_{i=1}^N$, where $\mathbf{x}_i^v$ denotes the v-th view sample of the i-th instance and d^v denotes the feature dimension in the v-th view, a multi-granularity self-learning guided deep multi-view clustering (MSG-MVC) method is proposed to partition these N samples into K disjoint clusters. MSG-MVC contains sample-guided information compression (SIC), representation-guided complementary integration (RCI), and cluster-guided consistent alignment (CCA), as shown in Fig. 1.

A. Sample-guided Information Compression

SIC aims to learn compact and informative representations of high-dimensional data in each view. To this end, SIC utilizes autoencoders to extract view-specific representations, and then leverages data as the self-learning signals to maximize the mutual information between view-specific representations and input data, which forces view-specific representations to preserve data as much as possible.

Specifically, given the v-th view data $\mathbf{X}^v = \{\mathbf{x}_i^v\}_{i=1}^N$, SIC obtains the view-specific representations $\mathbf{Z}^v$ via the v-th view encoder, i.e., $\mathbf{Z}^v = E^v(\mathbf{X}^v)$. Then, to preserve inherent data manifold, SIC maximizes the mutual information $I(\mathbf{X}^v; \mathbf{Z}^v)$ between $\mathbf{X}^v$ and $\mathbf{Z}^v$ as the sample-guided self-learning loss:

$$\mathcal{L}_{sic} = -\frac{1}{V}\sum_{v=1}^{V} I(\mathbf{X}^v; \mathbf{Z}^v) \tag{1}$$

The derivation process of the numerical solution of $I(\mathbf{X}^v; \mathbf{Z}^v)$ is as follows:

$$\begin{aligned} I(\mathbf{X}^v; \mathbf{Z}^v) &= \mathbb{E}_{p(\mathbf{x}^v, \mathbf{z}^v)}\left[\log \frac{p(\mathbf{x}^v, \mathbf{z}^v)}{p(\mathbf{x}^v)\, p(\mathbf{z}^v)}\right] \\ &= H(\mathbf{X}^v) + \mathbb{E}_{p(\mathbf{z}^v)}[\mathbb{E}_{p(\mathbf{x}^v|\mathbf{z}^v)}[\log p(\mathbf{x}^v|\mathbf{z}^v)]] \\ &\geq H(\mathbf{X}^v) + \mathbb{E}_{p(\mathbf{z}^v)}[\mathbb{E}_{p(\mathbf{x}^v|\mathbf{z}^v)}[\log D^v(\mathbf{x}^v|\mathbf{z}^v)]] \\ &= H(\mathbf{X}^v) + \mathbb{E}_{p(\mathbf{x}^v, \mathbf{z}^v)}[\log D^v(\mathbf{x}^v|\mathbf{z}^v)] \end{aligned} \tag{2}$$

where $D^v(\mathbf{X}^v|\mathbf{Z}^v)$ is a variational distribution approximating the true conditional distribution. In SIC, the decoder network is used to instantiate $D^v(\mathbf{X}^v|\mathbf{Z}^v)$.

Then, SIC introduces the view-specific clustering layer based on the heavy-tailed Student's t-distribution that can preserve manifold structure of data, to generate the cluster assignment $\mathbf{Q}^v \in \mathbb{R}^{N\times K}$:

$$q_{ij}^v = \frac{(1+\|\mathbf{z}_i^v - \mathbf{c}_j^v\|^2)^{-1}}{\sum_{k=1}^{K}(1+\|\mathbf{z}_i^v - \mathbf{c}_k^v\|^2)^{-1}} \tag{3}$$

where $\mathbf{C}^v = \{\mathbf{c}_1^v, \mathbf{c}_2^v, ..., \mathbf{c}_k^v\}$ is the view-specific learnable cluster prototypes initialized by K-means. $\mathbf{z}_i^v = \mathbf{Z}_{i,:}^v$ denotes the v-th view representation of $\mathbf{x}_i^v$ and q_{ij}^v is the probability that $\mathbf{z}_i^v$ belongs to the j-th cluster.

B. Representation-guided Complementary Integration

Current DMVC methods tend to aligning view-specific representations to aggregate complementary information, which results in fusion representations containing primarily consensus information, as shown in Fig. 2(a). Conversely, RCI leverages fusion representations as the self-learning signal to minimize the conditional entropy between fusion representations and view-specific representations, which endows fusion representations with rich complementary and consistent information across views, as shown in Fig. 2(b).

Given view-specific representations $[\mathbf{Z}^1, \mathbf{Z}^2, ..., \mathbf{Z}^V]$, to avoid subjective bias introduced by weighting, RCI utilizes the arithmetic mean fusion function to produce fusion representations $\mathbf{Z}$, i.e., $\mathbf{Z} = (\mathbf{Z}^1 + \mathbf{Z}^2 + ... + \mathbf{Z}^V)/V$, and then minimizes the conditional entropy $H(\mathbf{Z}^v|\mathbf{Z})$ as the representation-guided self-learning loss:

$$\mathcal{L}_{rci} = \frac{1}{V}\sum_{v=1}^{V} H(\mathbf{Z}^v|\mathbf{Z}) \tag{4}$$

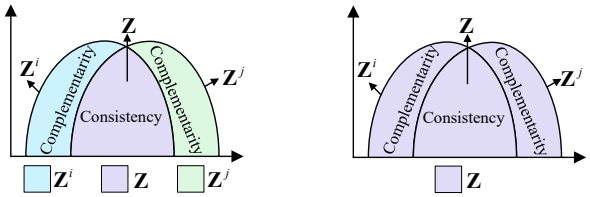
where $H(\mathbf{Z}^v|\mathbf{Z})$ expresses the uncertainty of $\mathbf{Z}^v$ under the given $\mathbf{Z}$ condition. $H(\mathbf{Z}^v|\mathbf{Z}) = 0$ indicates that fusion representations completely capture complementary information in the v-th view. Derivation of the solution of $H(\mathbf{Z}^v|\mathbf{Z})$ is :

$$\begin{aligned} H(\mathbf{Z}^v|\mathbf{Z}) &= -\mathbb{E}_{p(\mathbf{z}^v,\mathbf{z})}[\log p(\mathbf{z}^v|\mathbf{z})] \\ &= -\mathbb{E}_{p(\mathbf{z}^v,\mathbf{z})}[\log R(\mathbf{z}^v|\mathbf{z})] - D_{KL}(\mathcal{P}(\mathbf{Z}^v|\mathbf{Z})||R(\mathbf{Z}^v|\mathbf{Z})) \\ &\le -\mathbb{E}_{P(\mathbf{z}^v,\mathbf{z})}[\log R(\mathbf{z}^v|\mathbf{z})] \\ &= -\mathbb{E}_{P(\mathbf{z}^v,\mathbf{z})}\left[\log\left(\frac{1}{(2\pi\sigma^2)^{d_z^v/2}} e^{-\frac{\|\mathbf{z}^v - G^v(\mathbf{z})\|^2}{2\sigma^2}}\right)\right] \end{aligned} \tag{5}$$

where the variational distribution $R(\mathbf{Z}^v|\mathbf{Z})$ follows a Gaussian distribution $\mathcal{N}(\mathbf{Z}^v|G^v(\mathbf{Z}), \sigma^2\mathbf{I})$ is leveraged to approximate the true distribution $\mathcal{P}(\mathbf{Z}^v|\mathbf{Z})$. $\sigma^2\mathbf{I}$ is a variance matrix. $G^v(\cdot)$ is a self-learning transformation function from $\mathbf{Z}$ to $\mathbf{Z}^v$. For generality and simplicity, $G^v(\cdot)$ is implemented as an MLP.

C. Cluster-guided Consistent Alignment

CCA aims to alleviate the interference of noise hidden in rich complementary information to gain a robust clustering. To this end, CCA leverages cluster assignment results as the self-learning signals to enforce the multi-level semantic invariance across views for mining the consensus assignment.



(a) Consensus methods (b) MSG-MVC

Fig. 2. Different integrate strategies.

Given the fusion representation $\mathbf{Z}$, CCA utilizes K-means to obtain the common cluster prototypes $\mathbf{C}$, and then generates the common cluster assignment $\mathbf{Q}$ via the Student's t-distribution in Eq. (3). To sharpen semantic differences between clusters, the enhanced common cluster assignment $\mathbf{P} \in \mathbb{R}^{N\times K}$ is defined as:

$$p_{ij} = \frac{(q_{ij})^2}{\sum_{k=1}^{K}(q_{ik})^2} \tag{6}$$

where p_{ij} is probability of partitioning $\mathbf{z}_i$ to the j-th cluster.

Then, CCA maximizes the mutual information $I(\mathbf{P};\mathbf{Q}^v)$ between $\mathbf{P}$ and $\mathbf{Q}^v$ as the instance-level semantic-invariant self-learning loss $\mathcal{L}_{ins}$. Meanwhile, CCA treats each column in $\mathbf{P}$ and $\mathbf{Q}^v$ as cluster representations, and then maximizes the mutual information $I(\mathbf{P}^\top;(\mathbf{Q}^v)^\top)$ between cluster representations in $\mathbf{P}$ and $\mathbf{Q}^v$ belonging to the same cluster as the cluster-level semantic-invariant self-learning loss $\mathcal{L}_{clu}$.

$$\begin{aligned} \mathcal{L}_{cca} &= \mathcal{L}_{ins} + \mathcal{L}_{clu} \\ &= -\frac{1}{V}\sum_{v=1}^{V}(I(\mathbf{P};\mathbf{Q}^v) + I(\mathbf{P}^\top;(\mathbf{Q}^v)^\top)) \end{aligned} \tag{7}$$

Take $I(\mathbf{P};\mathbf{Q}^v)$ as an example to present the derivation process of the numerical solution of Eq. (7). Define $\mathbf{p}_i = \mathbf{P}_{i,:}$ and $\mathbf{q}_i^v = \mathbf{Q}_{i,:}^v$. Assuming that when $i \neq j$, $p(\mathbf{p}_i)$ and $p(\mathbf{q}_j^v)$ are independent and $p(\mathbf{p}_i, \mathbf{q}_i^v) > p(\mathbf{p}_i)p(\mathbf{q}_i^v)$ due to the correlation between positive pairs, and there exists a positive constant δ such that $p(\mathbf{q}_i^v|\mathbf{p}_i) > \delta$ for each sample. Based on $p(\mathbf{p}_i) \approx 1/N$ and $e^{sim(\mathbf{p}_i,\mathbf{q}_j^v)/\tau} \propto r_{ij}$ [13], it follows that:

$$\begin{aligned} I(\mathbf{P};\mathbf{Q}^v) &= \sum_{i=1}^{N}\sum_{j=1}^{N} p(\mathbf{p}_i,\mathbf{q}_j^v)\log r_{ii} \\ &\approx \sum_{i=1}^{N}\frac{1}{N}p(\mathbf{q}_i^v|\mathbf{p}_i)\log r_{ii} \\ &\ge \frac{\delta}{N}\sum_{i=1}^{N}\log\frac{r_{ii}}{\sum_{j=1}^{N} r_{ij}} + \frac{\delta}{N}\sum_{i=1}^{N}\log\sum_{j=1}^{N} r_{ij} \\ &\ge \frac{\delta}{N}\sum_{i=1}^{N}\log e^{sim(\mathbf{p}_i,\mathbf{q}_i^v)/\tau} \end{aligned} \tag{8}$$

TABLE I
THE STATISTICS OF THE MULTI-VIEW DATASETS.

Dataset	Samples	Views	Clusters	Dimensionality
BDGP	2500	2	5	1750/79
MSRC	210	5	7	24/576/512/256/254
Caltech-4V	1400	4	7	40/254/928/512
Caltech-5V	1400	5	7	40/254/1984/512/928
Scene-15	4485	3	15	20/59/40
ALOI-100	10800	4	100	77/13/64/125
YouTubeFace10	38654	4	10	944/576/512/640
AWA-7	10158	7	50	2688/2000/2000/2000/ 2000/4096/4096

where $r_{ij} = p(\mathbf{p}_i, \mathbf{q}_j^v)/(p(\mathbf{p}_i)p(\mathbf{q}_j^v))$. Similarly, the numerical solution of $I(\mathbf{P}^\top; (\mathbf{Q}^v)^\top)$:

$$I(\mathbf{P}^\top; (\mathbf{Q}^v)^\top) \geq \frac{1}{K}\sum_{k=1}^{K} \log \frac{e^{sim(\mathbf{p}_{:,k}, \mathbf{q}_{:,k}^v)/\tau}}{\sum_{j=1}^{K} e^{sim(\mathbf{p}_{:,k}, \mathbf{q}_{:,j}^v)/\tau}} \tag{9}$$

where $\mathbf{p}_{:,k} = \mathbf{P}_{:,k}$ and $\mathbf{q}_{:,k}^v = \mathbf{Q}_{:,k}^v$. The overall loss is:

$$\mathcal{L} = \mathcal{L}_{sic} + \alpha\mathcal{L}_{rci} + \beta\mathcal{L}_{cca} \tag{10}$$

where α and β are trade-off parameters. The final clustering result y_i is:

$$y_i = \arg\max_k \left(\frac{1}{V}\sum_{v=1}^{V} q_{ik}^v\right) \tag{11}$$

IV. EXPERIMENTS

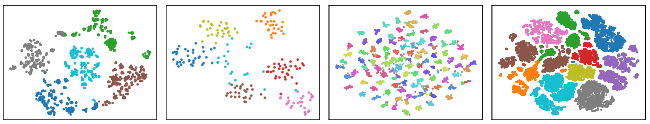
A. Experimental Settings

Datasets. BDGP [14], MSRC [15], Caltech-4V, Caltech-5V [16], Scene-15 [17], ALOI-100 [18], YTF-10 [19], and AWA-7 [20] are used as benchmarks. Table I lists the detailed information of the datasets.

Comparison methods. Nine state-of-the-art multi-view clustering methods, including SDMVC (2022 TKDE) [21], GCFAgg (2023 CVPR) [22] , MVCAN (2024 CVPR) [23], SCM (2024 IJCAI) [24], AEVC (2024 CVPR) [25], DMAC (2025 AAAI) [26], ALPC (2025 AAAI) [27], TLRLF (2025 TPAMI) [28], and PMIMC (2025 TIP) [29]. For a fair comparison, all comparison methods execute grid search experiment of hyperparameters in original papers for best performance.

Implementation details. We implement MSG-MVC in PyTorch and run experiments on an NVIDIA Tesla V100 GPU. The autoencoders are implemented by fully connected networks with the architecture D^v-500-500-2000-128-2000-500-500-D^v. The architecture of $G^v(\cdot)$ is 128-500-500-128. Pre-training for autoencoders and training are conducted for 200 and 500 epochs with a learning rate of 1e-4 and 1e-5, respectively. Batch size is 256 and the optimizer is Adam. The results are averaged of five runs. For BDGP, MSRC, Caltech-4V, Caltech-5V, and Scene-15, α is 0.01 and β is 0.001. For ALOI-100, YTF10, and AWA, α is 0.001 and β is 0.005.

Evaluation metrics. Evaluation metrics are accuracy (ACC), normalized mutual information (NMI), and purity (PUR). ACC measures clustering accuracy after optimal matching with ground-truth labels. NMI measures the normalized mutual dependence between predicted clusters and true labels. PUR measures cluster purity by assigning each cluster to its dominant ground-truth class and averaging over samples.



(a) BDGP (b) MSRC (c) ALOI-100 (d) YTF-10

Fig. 3. T-SNE visualization on four datasets.

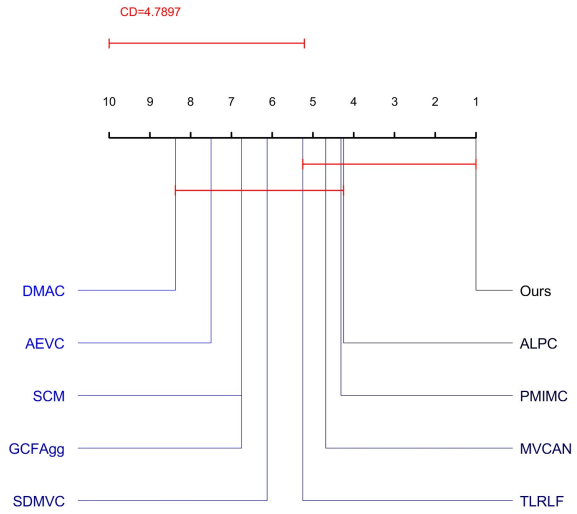


Fig. 4. The illustration of post-hoc Nemenyi test.

B. Performance Analysis

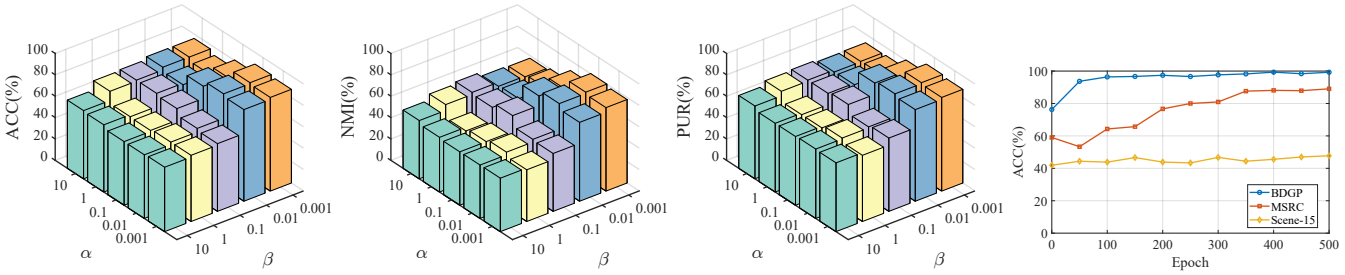
The results in Table II show that MSG-MVC achieves cutting-edge performance in DMVC. There are three observations. (1) MSG-MVC achieves the best performance on all eight datasets, demonstrating its overall effectiveness. (2) MSG-MVC achieves a significant performance advantage over the methods (GCFAgg, AEVC, ALPC, and TLRLF) that utilize inter-view invariance to aggregate information of views, which demonstrates the effectiveness of self-learning transformation via the conditional entropy minimization in fusing complementary information. (3) MSG-MVC outperforms the methods (SDMVC, MVCAN, DMAC, PMIMC) that only use pseudo labels to guide self-learning, which demonstrates the superiority and necessity of performing multi-granularity self-learning in different spaces for multi-view clustering. Fig. 3 shows the visualization results of t-SNE on four datasets, where MSG-MVC can produce compact clusters with clear inter-cluster boundaries.

To assess statistical significance, we conduct post-hoc Nemenyi tests on all datasets. The resulting critical-distance diagram in Fig.4 confirms that MSG-MVC is superior to most competing methods at the 5% significance level.

TABLE II
CLUSTERING RESULTS IN ACC%, NMI%, AND PUR%. THE BEST RESULTS ARE IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Methods	BDGP			MSRC			Caltech-4V			Caltech-5V		
	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
SDMVC(TKDE, 22)	<u>98.42</u>	<u>94.86</u>	<u>98.42</u>	77.62	70.24	80.00	80.14	66.62	80.14	83.86	73.92	83.86
GCFAgg(CVPR, 23)	98.08	94.01	98.08	71.90	62.77	75.24	76.43	<u>70.96</u>	76.43	83.93	74.25	83.93
MVCAN(CVPR, 24)	97.80	93.51	97.80	86.19	75.46	86.19	77.79	68.55	78.00	<u>87.64</u>	<u>78.39</u>	<u>87.64</u>
SCM(IJCAI, 24)	97.26	91.94	97.26	74.29	59.05	74.29	71.79	57.32	71.79	75.43	64.34	75.43
AEVC(CVPR, 24)	66.32	39.92	66.32	84.16	70.98	84.16	74.33	58.96	74.39	83.87	70.31	83.87
DMAC(AAAI, 25)	95.12	87.85	95.12	63.33	57.28	64.76	65.79	50.29	65.79	74.93	63.28	74.93
ALPC(AAAI, 25)	90.72	79.29	90.72	<u>88.57</u>	<u>78.45</u>	<u>88.57</u>	77.43	66.52	77.43	86.29	74.82	86.29
TLRLF(TPAMI, 25)	92.12	78.29	92.12	73.81	74.59	73.81	<u>80.86</u>	68.69	<u>80.86</u>	83.00	71.22	83.00
PMIMC(TIP, 25)	91.36	81.87	93.78	86.19	78.16	87.15	78.50	66.56	79.46	83.71	73.29	83.76
Ours	**99.28**	**97.46**	**99.28**	**89.05**	**80.18**	**89.05**	**83.93**	**71.58**	**83.93**	**90.86**	**83.52**	**90.86**

Methods	Scene-15			ALOI-100			YTF-10			AWA-7		
	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
SDMVC(TKDE, 22)	41.57	42.14	44.64	72.40	85.93	74.40	73.80	77.18	77.42	50.47	62.25	58.45
GCFAgg(CVPR, 23)	40.04	39.19	45.26	74.49	84.30	76.83	77.29	75.25	79.56	44.72	66.26	49.98
MVCAN(CVPR, 24)	43.86	<u>45.38</u>	48.99	76.26	<u>89.33</u>	78.89	76.29	80.14	80.36	50.98	66.41	59.30
SCM(IJCAI, 24)	39.78	36.75	41.18	<u>79.06</u>	87.76	<u>80.85</u>	81.40	78.99	82.58	47.58	56.12	52.16
AEVC(CVPR, 24)	41.40	40.83	45.60	70.12	84.05	72.86	74.41	81.43	79.39	55.40	63.88	61.66
DMAC(AAAI, 25)	41.92	42.29	45.98	72.19	84.23	75.17	77.72	80.53	82.80	30.42	39.74	35.32
ALPC(AAAI, 25)	43.88	45.15	<u>50.81</u>	77.98	88.78	70.98	83.19	81.62	86.14	56.21	60.40	62.90
TLRLF(TPAMI, 25)	<u>46.02</u>	45.08	50.03	74.33	82.35	76.96	80.81	78.82	83.05	57.62	66.79	63.23
PMIMC(TIP, 25)	45.28	42.69	46.86	74.69	86.54	75.01	<u>83.72</u>	<u>85.66</u>	<u>87.93</u>	<u>58.53</u>	<u>68.89</u>	<u>65.04</u>
Ours	**47.74**	**46.50**	**54.36**	**82.62**	**90.68**	**83.89**	**88.19**	**86.41**	**89.24**	**60.75**	**69.47**	**65.96**



(a) ACC (b) NMI (c) PUR (d) ACC vs. Epoch

Fig. 5. Parameter sensitivity analysis results. (a), (b), and (c) show the results on MSRC with respect to α and β. (d) shows the results on BDGP, MSRC, and Scene-15 under different numbers of training epochs.

TABLE III
ABLATION RESULTS IN ACC% AND NMI%. THE BEST RESULTS ARE IN BOLD.

Variants	BDGP		MSRC		ALOI	
	ACC	NMI	ACC	NMI	ACC	NMI
MSG-MVC w/o $\mathcal{L}_{sic}$	87.16	95.44	68.43	57.62	78.99	88.60
MSG-MVC w/o $\mathcal{L}_{rci}$	85.12	78.71	74.29	68.38	78.72	89.28
MSG-MVC w/o $\mathcal{L}_{cca}$	96.41	92.08	83.81	75.91	76.81	87.81
MSG-MVC w/o $\mathcal{L}_{ins}$	97.36	92.98	88.57	79.64	81.11	90.26
MSG-MVC w/o $\mathcal{L}_{clu}$	96.64	92.43	88.10	79.13	79.80	89.99
MSG-MVC	**99.28**	**97.46**	**89.05**	**80.18**	**82.62**	**90.68**

C. Ablation Study

Table III shows ablation results of each component in MSG-MVC. MSG-MVC w/o $\mathcal{L}_{sic}$, MSG-MVC w/o $\mathcal{L}_{rci}$, and MSG-MVC w/o $\mathcal{L}_{cca}$ denote the variants of MSG-MVC with $\mathcal{L}_{sic}$, $\mathcal{L}_{rci}$, and $\mathcal{L}_{cca}$ removed, respectively. MSG-MVC w/o $\mathcal{L}_{ins}$ and MSG-MVC w/o $\mathcal{L}_{clu}$ denote the variants where $\mathcal{L}_{ins}$ and $\mathcal{L}_{clu}$ are individually removed from $\mathcal{L}_{cca}$, respectively. There are three conclusions. (1) The five variants with different loss ablations produce inferior performance compared with MSG-MVC, validating the effectiveness of each loss. (2) The joint instance–cluster alignment outperforms either instance-level or cluster-level alignment alone, indicating that multi-level alignment yields more robust results. (3) MSG-MVC achieves the best result of ACC and NMI, validating the rationality and effectiveness of loss integration.

D. Parameter Sensitivity Analysis

To evaluate the sensitivity of MSG-MVC to the coefficients α and β, a hyper-parameter analysis is conducted on the MSRC dataset. Specifically, α and β are selected from $\{10, 1, 0.1, 0.01, 0.001\}$ where one parameter is fixed while the other is varied across the candidate set. The results in

Fig. 5(a)-(c) show that MSG-MVC is stable and achieves high performance when α and β are both small, particularly in $\{0.01, 0.001\}$.

To verify the effect of the number of training epochs on MSG-MVC, we record the clustering accuracy at different epochs on BDGP, MSRC, and Scene-15. The number of epochs is varied from 0 to 500 with a step size of 50. As shown in Fig. 5(d), MSG-MVC on MSRC improves steadily and stabilizes around 350 epochs. For BDGP, the performance improves rapidly in the early stage and becomes stable at around 100 epochs. On Scene-15, the performance increases gradually and stabilizes at around 350 epochs. Overall, the results on all three datasets become stable when the number of epochs reaches 350, indicating that the training process is sufficient and MSG-MVC has largely converged.

V. Conclusion

This paper proposes MSG-MVC, a multi-granularity self-learning guided deep multi-view clustering framework which leverages unsupervised signals at sample, representation, and semantic levels. Specifically, MSG-MVC introduces a sample-guided information compression module to capture intrinsic information within each view, a representation-guided complementary integration module that preserves cross-view complementarity via conditional-entropy minimization, and a cluster-guided consistent alignment module that enforces semantic consistency through joint instance-cluster alignment. Extensive experiments on eight benchmarks, compared with nine state-of-the-art methods, validate the effectiveness of MSG-MVC.

References

[1] T. Wang, F. Li, L. Zhu, J. Li, Z. Zhang, and H. T. Shen, "Cross-modal retrieval: a systematic review of methods and future directions," *Proceedings of the IEEE*, 2025.

[2] S. Jabeen, X. Li, M. S. Amin, O. Bourahla, S. Li, and A. Jabbar, "A review on methods and applications in multimodal deep learning," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 2s, pp. 1–41, 2023.

[3] S. Tang, Z. Zhang, Y. Zhang, J. Zhou, Y. Guo, S. Liu, S. Guo, Y.-F. Li, L. Ma, Y. Xue *et al.*, "A survey on automated driving system testing: Landscapes and trends," *ACM Transactions on Software Engineering and Methodology*, vol. 32, no. 5, pp. 1–62, 2023.

[4] J. Gao, M. Liu, P. Li, A. A. Laghari, A. R. Javed, N. Victor, and T. R. Gadekallu, "Deep incomplete multiview clustering via information bottleneck for pattern mining of data in extreme-environment iot," *IEEE Internet of Things Journal*, vol. 11, no. 16, pp. 26 700–26 712, 2024.

[5] P. Su, S. Huang, W. Ma, D. Xiong, and J. Lv, "Multi-view granular-ball contrastive clustering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 19, 2025, pp. 20 637–20 645.

[6] J. Gao, M. Liu, P. Li, J. Zhang, and Z. Chen, "Deep multiview adaptive clustering with semantic invariance," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 9, pp. 12 965–12 978, 2023.

[7] M. Zhao, W. Yang, and F. Nie, "Deep graph reconstruction for multi-view clustering," *Neural Networks*, vol. 168, pp. 560–568, 2023.

[8] Y. Zhang, J. Cai, Z. Wu, P. Wang, and S.-K. Ng, "Mixture of experts as representation learner for deep multi-view clustering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 21, 2025, pp. 22 704–22 713.

[9] Y. Gan, Y. You, J. Huang, S. Xiang, C. Tang, W. Hu, and S. An, "Multi-view clustering via multi-stage fusion," *IEEE Transactions on Multimedia*, 2025.

[10] Y. Lu, Q. Li, X. Zhang, and Q. Gao, "Deep contrastive representation learning for multi-modal clustering," *Neurocomputing*, vol. 581, p. 127523, 2024.

[11] Z. Lou, C. Zhang, H. Xue, Y. Ye, Q. Zhou, and S. Hu, "Self-supervised weighted information bottleneck for multi-view clustering," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024, pp. 4643–4650.

[12] H. Xu, Q. Wang, B. Wang, and Q. Gao, "Deep fair multi-view clustering with attention kan," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5061–5070.

[13] J. Xu, S. Chen, Y. Ren, X. Shi, H. Shen, G. Niu, and X. Zhu, "Self-weighted contrastive learning among multiple views for mitigating representation degeneration," *Advances in neural information processing systems*, vol. 36, pp. 1119–1131, 2023.

[14] X. Cai, H. Wang, H. Huang, and C. Ding, "Joint stage recognition and anatomical annotation of drosophila gene expression patterns," *Bioinformatics*, vol. 28, no. 12, pp. i16–i24, 2012.

[15] M.-S. Chen, L. Huang, C.-D. Wang, and D. Huang, "Multi-view clustering in latent embedding space," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 3513–3520.

[16] 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 *2004 conference on computer vision and pattern recognition workshop*. IEEE, 2004, pp. 178–178.

[17] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 2. IEEE, 2005, pp. 524–531.

[18] J.-M. Geusebroek, G. J. Burghouts, and A. W. Smeulders, "The amsterdam library of object images," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 103–112, 2005.

[19] J. Jin, S. Wang, Z. Dong, X. Liu, and E. Zhu, "Deep incomplete multi-view clustering with cross-view partial sample and prototype alignment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 11 600–11 609.

[20] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *International conference on machine learning*. PMLR, 2015, pp. 2152–2161.

[21] J. Xu, Y. Ren, H. Tang, Z. Yang, L. Pan, Y. Yang, X. Pu, P. S. Yu, and L. He, "Self-supervised discriminative feature learning for deep multi-view clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 7, pp. 7470–7482, 2022.

[22] 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, pp. 19 863–19 872.

[23] J. Xu, Y. Ren, X. Wang, L. Feng, Z. Zhang, G. Niu, and X. Zhu, "Investigating and mitigating the side effects of noisy views for self-supervised clustering algorithms in practical multi-view scenarios," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 22 957–22 966.

[24] C. Luo, J. Xu, Y. Ren, J. Ma, and X. Zhu, "Simple contrastive multi-view clustering with data-level fusion," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, vol. 5, 2024.

[25] S. Liu, K. Liang, Z. Dong, S. Wang, X. Yang, S. Zhou, E. Zhu, and X. Liu, "Learn from view correlation: An anchor enhancement strategy for multi-view clustering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 26 151–26 161.

[26] B. Wang, C. Zeng, M. Chen, and X. Li, "Towards learnable anchor for deep multi-view clustering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 20, 2025, pp. 21 044–21 052.

[27] Y. Chen, H. Wang, J. Peng, and Y. Wang, "Anchor learning with potential cluster constraints for multi-view clustering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 15, 2025, pp. 15 939–15 947.

[28] Z. Long, Q. Wang, Y. Ren, Y. Liu, and C. Zhu, "Tlrlf4mvc: Tensor low-rank and low-frequency for scalable multi-view clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[29] H. Yuan, Y. Sun, F. Zhou, J. Wen, S. Yuan, X. You, and Z. Ren, "Prototype matching learning for incomplete multi-view clustering," *IEEE Transactions on Image Processing*, 2025.