Deep Multi-View Contrastive Clustering via Graph Structure Awareness

Lunke Fei, *Senior Member, IEEE*, Junlin He, Qi Zhu, Shuping Zhao, *Member, IEEE*, Jie Wen, *Senior Member, IEEE*, and Yong Xu, *Senior Member, IEEE*

***Abstract*—Multi-view clustering (MVC) aims to exploit the latent relationships between heterogeneous samples in an unsupervised manner, which has served as a fundamental task in the unsupervised learning community and has drawn widespread attention. In this work, we propose a new deep multi-view contrastive clustering method via graph structure awareness (DMvCGSA) by conducting both instance-level and cluster-level contrastive learning to exploit the collaborative representations of multi-view samples. Unlike most existing deep multi-view clustering methods, which usually extract only the attribute features for multi-view representation, we first exploit the view-specific features while preserving the latent structural information between multi-view data via a GCN-embedded autoencoder, and further develop a similarity-guided instance-level contrastive learning scheme to make the view-specific features discriminative. Moreover, unlike existing methods that separately explore common information, which may not contribute to the clustering task, we employ cluster-level contrastive learning to explore the clustering-beneficial consistency information directly, resulting in improved and reliable performance for the final multi-view clustering task. Extensive experimental results on twelve benchmark datasets clearly demonstrate the encouraging effectiveness of the proposed method compared with the state-of-the-art models.**

***Index Terms*—Multi-view clustering, graph neural network, contrastive learning, self-supervision.**

I. Introduction

With the rapid development of multimedia, objects are usually described by diverse information from different views with large heterogeneous feature gaps. Multi-view clustering (MVC), which aims to partition unlabeled multi-view samples into different groups via similarity metric, has been a fundamental and pivotal task for various practical applications, such as recommendation systems, medical image analysis, and social network analysis [1], [2]. Over the past few decades, several methods have been proposed for MVC, including the subspace learning-based method [3], graph learning-based methods [4], and matrix factorization-based methods [5], [6]. For example, Wang et al. [7] applied an angle regulator based on subspace learning to enhance the multi-view consistent representation for MVC, and Nie et al. [8] proposed a self-weighted graph learning method for MVC by exploring the collaborative representation across different views. In addition, Zhao et al. [9] designed a deep multi-view concept learning (DMCL) method by hierarchically decomposing multi-view data via matrix factorization for clustering.

In recent years, owing to the powerful feature learning ability of deep neural networks, deep learning-based MVC methods have attracted increasing attention owing to their outstanding clustering performance, and the typical methods include deep-embedded clustering (DEC)-based [10], subspace clustering-based [11], [12], and GCN-based approaches [13]. The DEC-based methods apply view-specific encoders to learn the nonlinear properties of multi-view data for complex multi-view clustering. For example, Wen et al. [14] designed a cognitive deep incomplete multi-view clustering network (CDIMC-net) by combining graph embedding with view-specific deep encoders to capture the high-level features and local structure of samples for each view, and Wang et al. [15] designed a novel generative partial multi-view clustering network (GP-MVC) to explicitly generate missing data through a generative adversarial network for incomplete multi-view clustering. Unlike the learning attribute information of DEC-based methods, subspace clustering-based deep MVC methods aim to partition multi-view samples into different feature subspaces for MVC [16]. For example, Gao et al. [17] proposed a cross-modal clustering method that constrains the similarity of multi-modal data into different subspaces via deep canonical correlation analysis (DCCA), and Luo et al. [18] proposed a consistent and specific multi-view subspace clustering network to simultaneously exploit the view-specific and common representations of multi-view samples. In addition, the GCN-based MVC methods apply graph convolutional neural networks (GCNs) to exploit the structural relationships of multi-view data. For example, Cheng et al. [19] designed a multi-view attribute graph convolutional network (MAGCN) for clustering

Received 5 September 2024; revised 21 March 2025; accepted 19 May 2025. Date of publication 2 June 2025; date of current version 17 June 2025. This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFF0905602; and in part by the National Natural Science Foundation of China under Grant 62176066, Grant 62106052, and Grant 62372136. The associate editor coordinating the review of this article and approving it for publication was Dr. Zhengming Ding. *(Corresponding author: Jie Wen.)*

Lunke Fei, Junlin He, and Shuping Zhao are with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China (e-mail: flksxm@126.com; hejunlin200100@163.com; yb77458@connect.um.edu.mo).

Qi Zhu is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: zhuqi@nuaa.edu.cn).

Jie Wen is with Shenzhen Key Laboratory of Visual Object Detection and Recognition, Harbin Institute of Technology, Shenzhen, Shenzhen 518055, China (e-mail: jiewen_pr@126.com).

Yong Xu is with Shenzhen Key Laboratory of Visual Object Detection and Recognition, Harbin Institute of Technology, Shenzhen 518055, China, and also with the Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: yongxu@ymail.com).

Digital Object Identifier 10.1109/TIP.2025.3573501

1941-0042 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS. Downloaded on August 20,2026 at 14:42:30 UTC from IEEE Xplore. Restrictions apply.

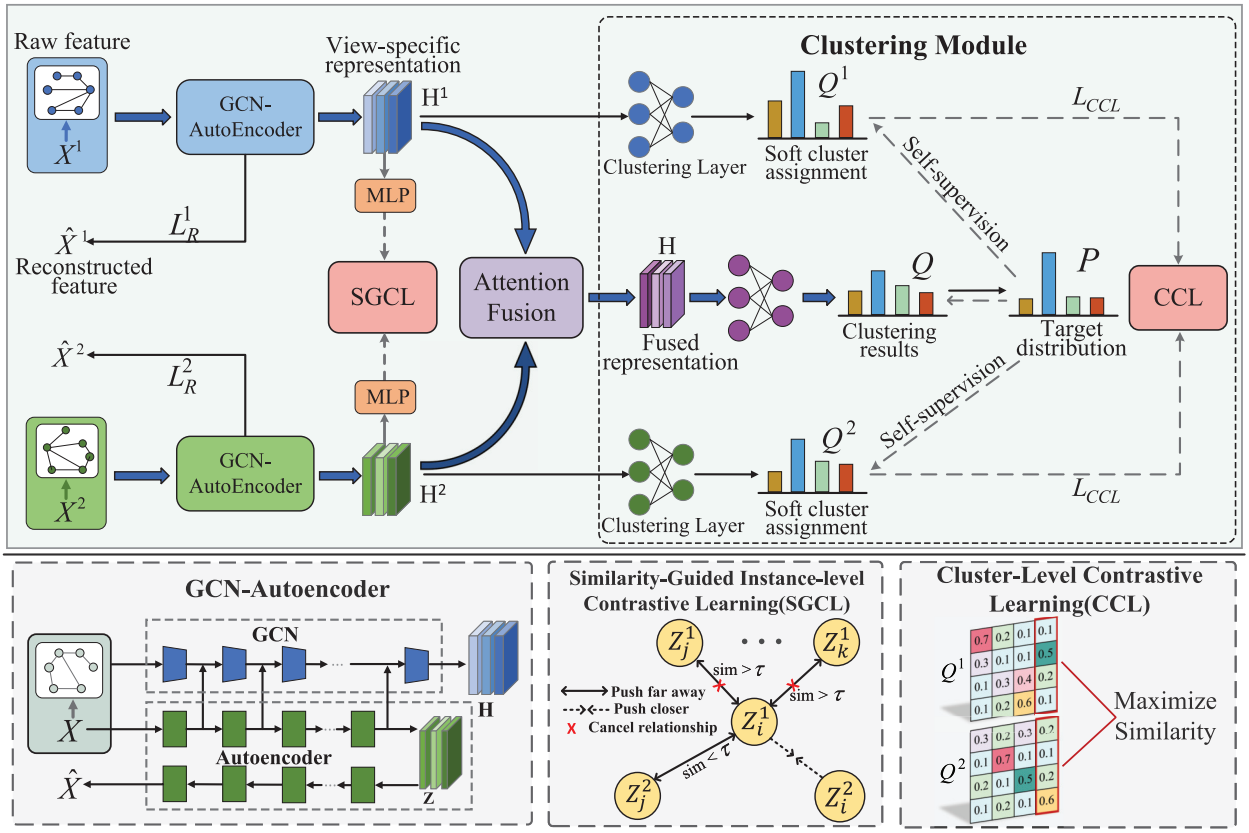


Fig. 1. Basic idea of the proposed DMvCGSA. First, we develop a GCN-AutoEncoder to exploit view-specific attribute features while preserving structural information. Then, we implement similarity-guided instance-level contrastive learning to align feature representations of different views, and further fuse view-specific features via an attention fusion module to obtain a more comprehensive representation. Finally, we perform contrastive learning at the clustering level and conduct a multiple self-supervised strategy to make the clustering results from different views consistent, improving the overall multi-view clustering performance.

by extracting graph structure relationships from multi-view data via a multi-view attribute graph convolutional encoder with an attention mechanism. While most existing deep MVC methods have achieved impressive clustering performance, they usually perform feature extraction from a single-source perspective, i.e., using either an autoencoder to extract attribute features or graph convolutional networks to extract structural information, making the extracted features less informative for modeling complex clustering structures. Additionally, these MVC methods for cross-view alignment through contrastive learning [20], [21] usually take the same instance from different views as positive pairs and the rest as negative pairs (including samples belonging to the same cluster), which possibly leads to the problem of false negative pairs in clustering scenarios, resulting in inaccurate clustering results.

In this paper, we propose a new deep multi-view contrastive clustering method via graph structure awareness (DMvCGSA). Fig. 1 shows the basic idea of the proposed method. Specifically, we first develop a GCN-AutoEncoder by combining the autoencoder with the GCN to exploit the attribute features while simultaneously preserving the high-order structural information. Then, we specifically design a similarity-guided instance-level contrastive learning scheme to pull within-cluster samples closer and push the between-cluster samples farther away in the contrastive feature space. Furthermore, we introduce an instance attention fusion strategy to adaptively incorporate the information across different views so that more complementary information can be exploited from multiple views. Finally, to make the view-specific clustering result consistent, we perform contrastive learning at the clustering level and employ a self-supervised strategy to guide the training of our model. By doing so, the clustering distributions from different views can be aligned, making the overall multi-view clustering result accurate and reliable. Extensive experimental results on twelve widely used datasets clearly show the promising effectiveness of the proposed method.

The main contributions of this work can be summarized as follows:

- We propose an end-to-end deep multi-view contrastive clustering method called DMvCGSA, which can extensively exploit the informative multi-view attribute features while simultaneously preserving the higher-order multi-view structural information, resulting in comprehensive and collaborative multi-view representations for improved clustering.
- We develop a similarity-guided instance-level contrastive learning scheme to specifically explore the relationships between samples within the same cluster, making the

Authorized licensed use limited to: NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS. Downloaded on August 20,2026 at 14:42:30 UTC from IEEE Xplore. Restrictions apply.

view-specific features discriminative, and further employ a cluster-level contrastive learning scheme to align the view-specific feature clustering results. By doing so, more informative multi-view representations can be obtained for accurate multi-view clustering.

- Compared with the state-of-the-art methods, our proposed method achieves superior clustering performance on twelve benchmark multi-view datasets.

The rest of this paper is organized as follows. In Section II, we briefly review the related topics. In Section III, we elaborate on the proposed DMvCGSA for multi-view clustering. In Section IV, we conduct extensive experiments to evaluate the proposed method. Finally, we conclude this paper in Section V.

II. Related Work

In this section, we first briefly review two related topics, namely, graph neural networks and contrastive learning, and then introduce the notation used in this paper.

A. Graph Convolutional Neural Networks

Graph convolutional networks (GCNs) [22] aim to capture the true insights of graph-structured data (such as molecular structures and knowledge graphs) via localized graph convolution operations, which have been widely used for diverse graph-based learning tasks, such as node classification and graph clustering [23]. For example, Kipf and Welling [24] proposed a GCN-based graph autoencoder (GAE) by integrating graph structures into node features for unsupervised representation learning, and Li et al. [25] applied a GCN-based adaptive graph autoencoder to adaptively learn an adjacency matrix for structural relationship representation. In recent years, several studies have applied GCNs to capture view-specific features for multi-view clustering. For example, Wang et al. [26] proposed a structure-adaptive unified graph neural network for multi-view clustering by embedding view-specific graph structures into a unified heterogeneous graph to form a consensus representation. In addition, Wang et al. [27] applied a GCN for incomplete multi-view clustering by inferring the missing features from GCN-based neighboring nodes. In this study, we incorporate the GCN into an autoencoder to extract multi-view attribute features while preserving the cross-view structural information for a more informative multi-view feature representation in MVC tasks.

B. Contrastive Learning

Contrastive learning has been widely used for various unsupervised learning tasks by maximizing the similarity of positive pairs and simultaneously minimizing that of negative pairs in contrastive feature space. SimCLR [28] is one of the most well-known methods for contrastive learning, which maximizes the similarity between augmented images and their original counterparts as positive pairs while minimizing the similarity of other pairs treated as negative samples to obtain more discriminative representations for self-supervised learning. Based on SimCLR, Li et al. [29] designed a contrastive learning network by performing instance- and cluster-level contrastive learning in the row and column spaces of the feature matrix for single-view clustering. Moreover, Xu et al. [20] conducted contrastive learning on both high-level features and semantic labels to uncover consistent information across different views for multi-view clustering, and Wang et al. [21] proposed a triple-granularity contrastive learning framework for deep multi-view subspace clustering. To overcome the false negative pairs in multi-view contrastive learning, Hu et al. [30] introduced cluster pseudo-labels into the construction of positive and negative pairs during contrastive learning for multi-view representation and clustering. However, it is noted that pseudo labels are hard to be correctly predicted at the early training stage. This possibly results in numerous wrong pseudo labels, making the learned feature less effective for multi-view clustering. In this study, we aim to use contrastive learning to explore the consistency of both the view-specific feature representations and view-specific clustering results for reliable multi-view clustering.

C. Notation

It is assumed that we have a multi-view dataset consisting of N samples with V views, represented as $X = \{X^1, X^2, \ldots, X^V\}$. The feature matrix in the v-th view is denoted as $X^v = \{x_1^v; x_2^v; \ldots, x_N^v\} \in R^{N \times d_v}$, where $x_i^v \in R^{d_v}$ denotes the feature vector of the i-th sample and d_v denotes the dimension of the feature vector in the v-th view. To obtain the graph structure relationship A^v of the multi-view data, a similarity relationship matrix, i.e., $S_{ij}^v = e^{-\frac{\|x_i^v - x_j^v\|_2^2}{2\sigma^2}} (i, j = 1, 2, \ldots, N)$, is usually constructed to measure the similarity between the x_i^v and x_j^v by calculating the Euclidean distance between two samples in each view. With the constructed similarity matrix S, the K-nearest neighbor algorithm can be used to construct an adjacent matrix, referred to as $A = \{A^1, \cdots, A^v, \cdots, A^V\}$, in which $A_{ij}^v \in R^{N \times N}$ indicates the existence of adjacent edges between the x_i^v and x_j^v.

III. Proposed Method

As shown in Fig. 1, our proposed DMvCGSA consists of a GCN-AutoEncoder, a similarity-guided instance-level contrastive learning module, an instance-level attention fusion module, and a clustering module. In this section, we introduce the proposed DMvCGSA with the above four submodules in detail.

A. GCN-AutoEncoder Module

In general, the attribute information describes the inherent characteristics of the samples, while the structural information characterizes the relationships among samples of different views [31]. In this study, we combine the autoencoder with GCN, named GCN-Autoencoder, to simultaneously exploit the attribute information and the affinities for multi-view samples. Specifically, we first encode the multi-view data in v-th the view into attribute feature descriptors by autoencoding as follows:

$$Z_{(l)}^v = \phi\left(Z_{(l-1)}^v W_{enc(l)}^v + b_{enc(l)}^v\right), \tag{1}$$

Authorized licensed use limited to: NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS. Downloaded on August 20,2026 at 14:42:30 UTC from IEEE Xplore. Restrictions apply.

where $W^v_{enc(l)}$ and $b^v_{enc(l)}$ represent the weight matrix and bias of the layer of the autoencoder in the v-th view, respectively. ϕ denotes the ReLU activation function. $Z^v_{(l)}$ represents the attribute features learned by the l-th layer of the autoencoder, and $Z^v_{(0)}$ denotes the original samples in the v-th view, i.e., X^v.

While attribute feature encoding is performed, we simultaneously employ a multi-layer GCN to capture the structural information of multi-view data, which can be expressed as follows:

$$H^v_{(l)} = f\left(\tilde{D}^{v-\frac{1}{2}}\tilde{A}^v\tilde{D}^{v-\frac{1}{2}}\tilde{H}^v_{(l-1)}W^v_{(l)} + b^v_{(l)}\right), \tag{2}$$

where $\tilde{A}^v = A^v + I_n \in R^{N\times N}$ and I_n is an identity matrix. $\tilde{D}^v_{ii} = \sum_j \tilde{A}^v_{ij}$ represents the degree matrix of $\tilde{A}^v$. $f(\cdot)$ represents the graph convolution operation of the GCN. In addition, $W^v_{(l)}$ and $b^v_{(l)}$ denote the learnable parameter matrix and the bias of the l-th graph convolutional layer in the v-th view, respectively. $H^v_{(l)}$ presents the feature representation obtained by the GCN in the l-th layer, and $H^v_{(0)} = X^v$. After that, we iteratively fuse the structural information (i.e., $H^v_{(l-1)}$) with the attribute feature (i.e., $Z^v_{(l-1)}$) in a layer-by-layer manner, as follows:

$$\tilde{H}^v_{(l-1)} = (1-\mu)\,H^v_{(l-1)} + \mu Z^v_{(l-1)}, \tag{3}$$

where $\tilde{H}^v_{(l-1)}$ represents the combined representation of the structural and attribute information. μ is a balance coefficient used to control the weights of the structural information and attribute features. Notably, $\tilde{H}^v_{(l-1)}$ is used as the input of the l-th layer of the GCN to generate the $H^v_{(l)}$. In this way, we can capture the high-order structural information with simultaneously preserving the inherent attribute information of the multi-view samples, such that more informative multi-view representation can be obtained.

To reduce the information loss during autoencoding, we embed a decoder into the autoencoder to recover the attribute features (i.e., $Z^v_{(l)}$) of the multi-view data representation, referred to as $\hat{X}^v$, and minimize the reconstruction error by calculating the Frobenius norm between the X^v and $\hat{X}^v$, as follows:

$$L_R = \sum_{v=1}^{V}\left\|X^v - \hat{X}^v\right\|_F^2. \tag{4}$$

B. Similarity-Guided Instance-Level Contrastive Learning

To align data from different views consistently while enhancing the discriminative power of feature representations, contrastive learning has been widely employed to simultaneously maximize the distance between positive pairs and minimize the distance between negative pairs [28]. However, most existing contrastive learning methods [20], [21] directly use the same instance across different views as positive pairs and the others as negative sample pairs. This easily leads to large false negative pairs because the sample pairs have the same clustering labels but far feature distances.

In general, within-cluster samples usually exhibit high similarity, whereas inter-cluster samples display low similarity. Inspired by this, we define a similarity threshold t, and the negative sample pairs with greater similarities than this threshold are considered false negative pairs. Subsequently, we remove these potential false negative pairs from the negative sample set, which are not added into the positive sample set due to their uncertainty as true positive samples. To be more specific, we first map the view-specific feature (i.e.,H^v) into the contrastive space representation via two fully connected layers, as follows:

$$F^v_i = MLP\left(H^v_i\right), \tag{5}$$

where F^v_i represents the embedded feature map of the i-th sample. Then, we measure the sample similarity in the contrastive space via the cosine metric as follows:

$$s\left(F^m_i, F^n_j\right) = \frac{\left(F^m_i\right)\left(F^n_j\right)^T}{\left\|F^m_i\right\|\cdot\left\|F^n_j\right\|}. \tag{6}$$

Afterward, the loss between two different views of the same instance can be calculated as follows:

$$\begin{aligned} l^{mn}_i &= -\log\frac{\exp\left(s\left(F^m_i, F^n_i\right)/\tau_I\right)}{M+U},\\ M &= \sum_{j=1}^{N}\left[I_{\left[s\left(F^m_i,F^m_j\right)<t\right]}\exp\left(s\left(F^m_i, F^m_j\right)/\tau_I\right)\right],\\ U &= \sum_{j=1}^{N}\left[I_{\left[s\left(F^m_i,F^n_j\right)<t\right]}\exp\left(s\left(F^m_i, F^n_j\right)/\tau_I\right)\right], \end{aligned} \tag{7}$$

where τ_I is the instance-level temperature parameter [28]. $I_{\left[s\left(F^m_i,F^n_j\right)<t\right]} \in \{0,1\}$ is an indicator function, that equals 1 when $s\left(F^m_i, F^n_j\right) < t$, and 0 otherwise. Furthermore, the total similarity-guided instance-level contrastive loss for all samples in multiple views can be computed as follows:

$$L_{SGCL} = \frac{1}{2N}\sum_{m=1}^{V}\sum_{n=1}^{V}\sum_{i=1}^{N} I_{[m\neq n]}l^{mn}_i. \tag{8}$$

By minimizing L_{SGCL} during contrastive learning, we align the representations across different views while ensuring that only the distances between sample pairs with low similarity are minimized in contrastive feature space. This prevents the distances between samples that may belong to the same cluster from increasing during the alignment process.

C. Instance-Level Attention Fusion

Different views of an instance not only share the consistency information but also convey the view-specific characteristics of the instance. To extensively exploit the complementary features of multiple views, we adaptively fuse view-specific features by calculating the attention weights of each instance at different views. To be more specific, we first calculate the intermediate attention weights of the multi-view features, referred to as $B \in R^{N\times V}$, via a feedforward network (FFN), as follows:

$$B = FFN\left(\text{concat}\left(H^1, H^2, \ldots, H^V\right)\right), \tag{9}$$

where *concat* concatenates the multi-view features in a horizontal manner. After that, we calculate the attention coefficients of all instances in each view as follows:

$$[e^1, e^2, \ldots, e^V] = \text{softmax}\left(\frac{\text{sigmoid}(B)}{\tau}\right), \tag{10}$$

Authorized licensed use limited to: NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS. Downloaded on August 20,2026 at 14:42:30 UTC from IEEE Xplore. Restrictions apply.

where e^V represents the attention coefficient of instances in the v-thview. τ is a temperature parameter that controls the sharpness of e^v. In addition, the sigmoid function operates on the intermediate attention weights to avoid an excessively high score assigned to a view, such that a balanced weight distribution can be obtained. Finally, we apply an attention fusion scheme to obtain the consensus representation of multiple views as follows:

$$H = \sum_{v=1}^{V} \lambda^v \odot H^v, \tag{11}$$

where $\lambda^v = [e^v, e^v, \ldots, e^v] \in R^{N \times d_v} (v = 1, 2, \ldots, V)$ and $\odot$ represents the matrix element-wise multiplication. By adaptively incorporating different views, complementary information from multi-view can be conveyed in the final global feature descriptor H.

D. Clustering Module

The clustering module includes a cluster-level contrastive learning module and a multiple self-supervised mechanism. To make the view-specific clustering consistent with the final multi-view clustering results, we introduce cluster-level contrastive learning to exploit the correlation between the clustering results of different views. First, we conduct view-specific clustering as follows:

$$Q^v = g(H^v), \quad v = 1, 2, \ldots, V, \tag{12}$$

where $g(\cdot)$ denotes the view-parameter-shared clustering layer, including two fully connected layers followed by a Softmax function. $Q^v \in R^{N \times C}$ represents the cluster assignments for the v-th view, and C is the total number of clusters. $Q^v_{:,j}$ is the j-th column of Q^v, which denotes the probability values for all the instances assigned to the j-th cluster in the v-th view.

Inspired by the fact that different views of an instance usually share similar semantic information, we treat the assignment of the same cluster in different views (e.g., $Q^m_{:,j}$and $Q^n_{:,j}$) as positive pairs, and the others as negative pairs. Similar to instance-level contrastive learning, we calculate the cluster-level contrastive loss as follows:

$$\begin{aligned} \hat{l}^{mn}_j &= -\log \frac{\exp\left(s\left(Q^m_{:,j}, Q^n_{:,j}\right)/\tau_C\right)}{M' + U'}, \\ M' &= \sum_{c=1}^{C} \exp\left(s\left(Q^m_{:,c}, Q^m_{:,j}\right)/\tau_C\right), \\ U' &= \sum_{c=1}^{C} \exp\left(s\left(Q^m_{:,c}, Q^n_{:,j}\right)/\tau_C\right), \end{aligned} \tag{13}$$

where $Q^m_{:,j}$ and $Q^n_{:,j}$ represent the cluster assignment probabilities of the j-th cluster in the m-th and n-th views, respectively. τ_C is the cluster-level temperature parameter [29]. After that, we can obtain the total cluster-level contrastive learning loss as follows:

$$L_{CCL} = \frac{1}{2C} \sum_{m=1}^{V} \sum_{n=1}^{V} \sum_{c=1}^{C} \left[I_{[m \neq n]} \hat{l}^{mn}_j - H(Q) \right], \tag{14}$$

where $H(Q) = \sum_{j=1}^{C} \left[\sum_{v=1}^{V} G\left(Q^v_{:,j}\right) \log G\left(Q^v_{:,j}\right) \right]$represents the entropy of clustering assignment probabilities and where $G\left(Q^v_{:,j}\right) = \left(\sum_{i=1}^{N} Q^v_{ij}\right)/N$, which is used to prevent a trivial solution [32]. By minimizing the cluster-level contrastive loss L_{CCL}, we can extensively leverage the consistency information among the clustering results of different views, resulting in improved clustering performance.

Moreover, we introduce a multiple self-supervised strategy to ensure the consistency of cross-view cluster assignments. To increase the discriminability of the soft clustering assignments during $Q = g(H) \in R^{N \times C}$, we calculate the unified auxiliary target distribution $P \in R^{N \times C}$ [33] as follows:

$$P_{ij} = \frac{Q^2_{ij}}{\sum_{j=1}^{C} Q^2_{ij}}. \tag{15}$$

After that, we design a multiple self-supervised loss based on KL-divergence as follows to train the clustering network in a self-supervised manner:

$$\begin{aligned} L_{KL} &= KL\left(\frac{Q + \sum_{v=1}^{V} Q^v}{(V+1)} \| P\right) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{C} P_{ij} \log \frac{P_{ij}}{\left(Q_{ij} + \sum_{v=1}^{V} Q^v_{ij}\right)/(V+1)}. \end{aligned} \tag{16}$$

By minimizing the L_{KL}, we can align the clustering assignments (i.e., Q^v and Q) from multiple views with the high-confidence auxiliary target distribution (i.e., P). As a result, during multi-view clustering, our proposed method can adaptively achieve consistent clustering results at different views.

E. Objective Function

By combining the reconstruction error, similarity-guided instance-level contrastive loss, cluster-level contrastive loss, and multiple self-supervised loss (i.e., L_R, L_{HSCL}, L_{CCL} and L_{KL}), our proposed method derives the overall objective loss function as follows:

$$L = L_R + \lambda_1 L_{SGCL} + \lambda_2 L_{CCL} + \lambda_3 L_{KL}, \tag{17}$$

where λ_1, λ_2 and λ_3 are the trade-off hyperparameters. Algorithm 1 summarizes the optimization process of the proposed network, which can be trained in an end-to-end manner. The final clustering result $Y \in R^N$ is obtained from the clustering prediction probability Q, i.e., $Y_i = \arg\max_{j \in \{1,2,\ldots,C\}} Q_{ij}$.

IV. Experiments

We evaluate our proposed method on twelve public multi-view datasets and compare it with ten state-of-the-art methods. All the experiments are conducted with Ubuntu 23.04 run on a 24 GB NVIDIA GeForce RTX 3090 Linux server. For our

Authorized licensed use limited to: NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS. Downloaded on August 20,2026 at 14:42:30 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1 DMvCGSA

input: Dataset $X = \{X^1, X^2, ..., X^V\}$, number of clusters C, number of neighbors K, threshold t, balance coefficient μ, trade-off parameters λ_1, λ_2, and λ_3, maximum number of iterations $epochs$.

output: Clustering predictions Q

Calculate the adjacency matrix $\{A^v\}_{v=1}^V$;
for $epoch = 1$ to $epochs$ **do**
  Compute the view-specific representations $\{H^v\}_{v=1}^V$ via Eq. (2);
  Calculate the loss L_R via Eq. (4);
  Calculate the similarity-guided instance-level contrastive loss L_{SGCL} via Eq. (8);
  Compute fused representations H via Eq. (11);
  Compute the soft assignment for each view $Q^v = g(H^v)$ and the global view $Q = g(H)$;
  Calculate the cluster-level contrastive loss L_{CCL} via Eq. (14);
  Calculate the target distribution P via Eq. (15);
  Calculate the multiple self-supervised loss L_{KL} via Eq. (16);
  Calculate the total loss L via Eq. (17);
  Update L through the gradient descent learning scheme;
end for
return Q

TABLE I
SUMMARY OF THE MULTI-VIEW DATASETS

Dataset	Samples	Category	Views	Feature Dimension
Scene-15	4,485	15	2	59/20
MSRC-V1	210	7	2	576/512
BDGP	2,500	5	2	1750/79
Reuters	18,758	6	2	10/10
Caltech-2V	1,400	7	2	40/254
Wiki	2,866	10	2	128/10
HandWritten	2,000	10	2	76/64
VOC	5,649	20	2	512/399
Caltech-3V	1,400	7	3	40/254/928
Fashion	10,000	10	3	784/784/784
Prokaryotic	551	4	3	438/3/393
MNIST-10K	10,000	10	3	30/9/30

proposed method, we implement the DMvCGSA via PyTorch 2.1.2 and use the Adam optimizer with the default learning rate of 10^{-3}. In addition, the temperature parameters τ, τ_I and τ_C are empirically set to 10, 1.0, and 0.5, respectively. The kernel bandwidth of a Gaussian function is fixed at σ = 1.0. Throughout the experiment, the encoder part of the autoencoder consists of five fully connected layers, and the GCN is composed of five graph convolutional layers. The dimensions of each layer are input-1024-1024-1024-output. The input dimension is always the feature dimension of the original data, and the output dimension is fixed at 512.

A. Databases

We employ twelve widely used datasets to conduct our comparative experiments, including the Scene-15, MSRC-V1, BDGP, Reuters, Caltech-2V, Wiki, HandWritten, VOC, Caltech-3V, Fashion, and Prokaryotic datasets, as summarized in Table I.

The **Scene-15** [34] dataset contains 4,485 images collected from fifteen categories of indoor and outdoor scenes. Following [35], we extract the 20-dimensional global image scene template (GIST) features and the 59-dimensional pyramid histogram of oriented gradient (PHOG) features as two different scene views for our experiments.

The **MSRC-V1** [36] dataset consists of 210 scene images captured from seven categories, including trees, buildings, airplanes, cows, faces, cars, and bicycles, and the GIST features and histogram of oriented gradient features (HOG) are extracted as two views of the scene images.

The **BDGP** [37] dataset contains 2,500 Drosophila embryo images across five categories. Similar to [20], we extract the 1750-dimensional SIFT features and 79-dimensional gene ontology textual features as the two views for our experiments.

The **Reuters** [38] dataset comprises 18,758 newswire articles collected from six categories, and each article is represented in five languages such as English, French, German, Italian, and Spanish. Following [39], we project the languages English and French articles into 10-dimensional space, using these projections as the two views of the articles for our experiments.

The **Caltech-2V** and **Caltech-3V** [40] datasets contain 1,400 RGB images collected from seven different categories, where the Caltech-2V employs the 40-dimensional wavelet moment (WM) and 254-dimensional census transform histogram (CENTRIST) representations as the two views of the images, and the Caltech-3V extends the Caltech-2V by adding 928-dimensional local binary pattern (LBP) features as the third view.

The **Wiki** [41] dataset consists of 2,866 images across ten categories. In this experiment, we extract the 128-dimensional SIFT feature and 10-dimensional latent Dirichlet allocation (LDA)-based feature vectors as the two-view representations of the samples.

The **HandWritten** [42] dataset includes 2,000 digital images of 0 to 9 handwritten characters, and each digital image contains 200 samples. Similar to [35], we utilize the 76-dimensional Fourier coefficient of the character shapes and the 64-dimensional Karhunen-Love coefficient as the two views of the digital images to conduct our experiments.

The **VOC** [43] dataset consists of twenty clusters of 5,649 image-text pairs, where the images and text are considered as the two views of the sample pairs, represented as the GIST feature descriptors and word frequency vectors, respectively.

The **Fashion** [44] dataset is an image dataset of about fashion products (T-shirts, coats, etc.), which contains 10,000 samples from a total of ten categories. Similar to [45], we take pictures of three different design styles of a product as three views.

The **Prokaryotic** [46] dataset consists of 551 prokaryotic samples across 4 categories. Each sample is characterized by a textual view and two genomic views. The textual view is a document term matrix that records the TF-IDF [47] reweighted word frequencies and the genomic views include the proteomic composition and gene pool.

Authorized licensed use limited to: NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS. Downloaded on August 20,2026 at 14:42:30 UTC from IEEE Xplore. Restrictions apply.

TABLE II
ACC, NMI AND ARI OBTAINED BY DIFFERENT METHODS ON EIGHT TWO-VIEW DATABASES, WHERE THE BEST AND SECOND-BEST RESULTS ARE BOLDED AND UNDERLINED, RESPECTIVELY

DataSet	Scene-15			MSRC-V1			BDGP			Reuters		
Evaluation Metrics	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
BSV	29.77	26.64	18.08	69.52	55.19	47.31	56.28	44.40	15.14	47.97	16.07	12.14
Concat	33.48	28.77	13.25	80.47	68.67	62.86	60.92	48.34	19.21	43.42	22.79	16.05
DEMVC (21, Inf.Sci.)	34.60	41.74	21.97	72.38	63.40	55.74	91.80	82.40	80.95	45.25	32.23	24.57
SiMVC (21, CVPR)	30.68	27.54	13.55	69.52	63.58	52.88	49.76	29.06	22.72	32.52	7.55	6.86
CoMVC (21, CVPR)	32.64	31.57	16.77	71.90	64.74	56.32	50.30	30.88	24.15	39.19	16.75	15.20
MFLVC (22, CVPR)	36.30	33.80	21.80	66.67	61.14	48.50	98.36	95.10	95.98	45.16	**38.14**	25.26
DealMVC (23, MM)	30.17	25.97	16.29	68.57	65.03	54.87	96.24	91.94	91.08	49.49	31.83	21.97
CVCL (23, ICCV)	39.31	39.38	22.88	89.90	82.01	79.12	98.20	94.15	95.58	48.50	32.14	25.83
SEM (23, NeurIPS)	38.11	40.21	21.58	62.86	53.80	43.11	85.28	73.93	64.08	45.75	27.55	18.66
ICMVC (24, AAAI)	39.73	37.41	22.53	87.24	78.18	74.78	97.72	92.63	94.41	49.83	32.89	24.99
OURS	**41.82**	**42.53**	**24.35**	**90.48**	**82.44**	**79.59**	**98.60**	**95.14**	**96.54**	**52.79**	35.40	**27.36**
DataSet	Caltech-2V			Wiki			HandWritten			VOC		
Evaluation Metrics	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
BSV	47.10	31.09	24.67	19.50	3.88	1.61	73.40	68.91	57.60	30.58	15.31	6.50
Concat	49.71	33.45	25.50	21.75	13.43	40.04	77.10	73.41	63.43	40.25	35.91	11.80
DEMVC (21, Inf.Sci.)	34.21	23.18	12.50	30.11	24.57	16.79	70.30	70.64	56.16	32.18	31.35	19.51
SiMVC (21, CVPR)	47.28	32.44	25.64	51.46	50.36	38.23	72.85	70.09	60.57	55.54	57.86	43.35
CoMVC (21, CVPR)	59.92	43.00	36.56	58.12	54.08	42.81	66.15	69.43	55.68	51.77	57.54	41.73
MFLVC (22, CVPR)	61.07	52.63	41.93	48.53	35.92	28.59	84.30	80.71	75.66	31.32	27.15	15.70
DealMVC (23, MM)	59.79	50.16	39.36	55.99	48.94	40.13	82.75	79.35	74.36	40.27	34.94	25.44
CVCL (23, ICCV)	63.29	**53.21**	43.12	48.46	44.13	33.06	77.70	79.47	70.87	51.34	51.47	40.18
SEM (23, NeurIPS)	60.79	52.49	40.08	51.85	50.06	37.98	82.90	80.33	73.53	36.77	31.24	19.23
ICMVC (24, AAAI)	20.97	4.07	1.75	49.13	40.09	32.36	83.48	82.16	75.91	45.25	50.58	33.56
OURS	**63.52**	52.28	**43.47**	**61.13**	**54.40**	**45.06**	**85.08**	**84.44**	**78.32**	**57.04**	**59.61**	**50.97**

The **MNIST-10K** [48] dataset consists of 10,000 images of handwritten numbers from 0-9. We use 30-dimensional IsoProjection, 9-dimensional linear discriminant analysis (LDA), and 30-dimensional neighborhood preserving embedding (NPE) as three different views to perform our experiments.

B. Two-View Clustering Results

Notably, two views are basic multi-view scenarios and are most widely used for object description [35]. In this subsection, we conduct two-view clustering experiments to evaluate our proposed method on eight two-view databases, including the Scene-15, MSRC-V1, BDGP, Reuters, Caltech-2V, Wiki, HandWritten, and VOC datasets. For a better evaluation, we compare our proposed method with state-of-the-art methods, including typical single-view clustering methods, such as BSV [49] and Concat [35], and representative deep multi-view clustering methods, including the DEMVC [50], SiMVC [51], CoMVC [51], MFLVC [20], DealMVC [52], CVCL [53], SEM [54] and ICMVC [35] methods. Following the standard evaluation protocol [33], we calculate three widely used metrics to evaluate clustering performance: accuracy (ACC), normalized mutual information (NMI), and the adjusted Rand index (ARI). Following [35], for the single-view-based BSV method, we perform k-means clustering for each view and report the best ones as the results. For the Concat method, we first concatenate features from all views into a common representation and then perform k-means clustering. All the comparison methods are repeated five times, and the average evaluation results are reported, as tabulated in Table II.

Table II shows that our proposed method consistently outperforms most state-of-the-art methods except on the Fashion dataset. Specifically, our proposed method achieves obviously higher ACC, NMI and ARI than the two single-view-based BSV and Concat methods, possibly because fewer views of information are used for the single-view methods. Moreover, our proposed method consistently outperforms deep multi-view clustering methods such as SiMVC, CoMVC and SEM. This may be because SiMVC, CoMVC and SEM consider only instance-level contrastive learning, whereas our proposed DMvCGSA introduces a similarity-guided instance-level contrastive learning strategy to explore between-view consistency information, such that false negatives can be alleviated during multi-view clustering information learning and better clustering performance can be achieved. In particular, our method outperforms the GCN-based ICMVC and the autoencoder-based MFLVC, DealMVC and CVCL methods on almost all the datasets. The reason is that our proposed DMvCGSA can specifically learn both the structural information and attribute features of multi-view information, such that the underlying structure can be accurately characterized, resulting in improved multi-view clustering accuracy.

We further visualize the t-SNE [55] based on the feature maps obtained by the proposed DMvCGSA method with different training epochs, as shown in Fig. 2. Our method can achieve compact multi-view clustering results with a sufficient number of learning steps, demonstrating the promising effectiveness of the proposed method.

C. Three-View Clustering Results

To further evaluate our proposed method on multi-view clustering with more than two views, we conduct three-view clustering experiments on four three-view datasets, including the Caltech-3V, Fashion, Prokaryotic and MNIST-10K datasets, as reported in Table III. We can see that our proposed

Authorized licensed use limited to: NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS. Downloaded on August 20,2026 at 14:42:30 UTC from IEEE Xplore. Restrictions apply.

TABLE III

ACC, NMI AND ARI OBTAINED VIA DIFFERENT METHODS ON THE THREE-VIEW CALTECH-3V, FASHION, PROKARYOTIC AND MNIST-10K DATASETS

DataSet	Caltech-3V			Fashion			Prokaryotic			MNIST-10K		
Evaluation Metrics	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
BSV	46.85	31.33	24.56	56.97	51.01	35.25	30.62	9.33	5.80	88.63	76.76	76.72
Concat	49.21	33.43	25.32	67.25	69.87	54.67	36.44	11.07	5.99	59.54	53.58	41.46
DEMVC (21, Inf.Sci.)	40.00	28.81	20.58	66.13	80.28	64.03	52.81	27.11	18.48	55.98	44.43	43.02
SiMVC (21, CVPR)	54.09	48.66	39.89	59.80	60.89	44.72	39.20	10.23	6.45	66.59	62.06	50.62
CoMVC (21, CVPR)	52.57	40.02	32.32	78.33	77.23	67.68	49.18	27.60	**21.53**	84.39	74.59	70.83
MFLVC (22, CVPR)	63.29	55.01	47.11	99.07	97.78	97.96	41.74	22.81	10.33	90.47	80.28	80.28
DealMVC (23, MM)	58.36	54.26	41.62	90.05	94.52	88.69	53.36	28.28	11.43	71.50	66.92	59.53
CVCL (23, ICCV)	62.14	52.62	43.74	99.13	97.82	98.09	52.00	23.72	12.75	90.96	81.86	81.38
SEM (23, NeurIPS)	64.02	54.34	45.68	**99.33**	**98.23**	**98.53**	48.82	21.58	11.83	89.40	77.85	78.30
ICMVC (24, AAAI)	21.45	4.10	2.76	92.87	87.39	85.82	48.10	23.65	13.44	91.90	82.69	83.30
OURS	**66.19**	**55.77**	**47.19**	97.59	94.76	94.86	**54.63**	**28.50**	17.78	**92.53**	**83.75**	**84.52**

TABLE IV

CLUSTERING RESULTS OBTAINED BY DIFFERENT VARIANTS OF THE PROPOSED METHODS BY REMOVING DIFFERENT COMPONENTS FROM THE DMVCGSA FRAMEWORK ON EIGHT MULTI-VIEW DATASETS

Methods	Scene-15			MSRC-V1			BDGP			Reuters		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
(1) L_{SGCL}	35.94	40.47	18.70	84.76	77.44	65.81	80.79	74.44	64.04	50.45	30.67	20.32
(2) L_{CCL}	39.66	40.62	22.90	89.37	80.69	77.80	76.73	68.62	62.60	47.99	29.62	24.07
(3) L_{KL}	34.30	36.35	18.19	76.35	68.66	58.77	30.72	24.12	6.12	46.37	25.40	22.89
(4) L_{SGCL} + L_{CCL}	40.17	42.01	23.58	90.00	81.41	78.80	92.75	88.35	87.66	51.76	31.88	26.36
(5) L_{SGCL} + L_{KL}	36.23	41.00	19.46	86.19	78.28	69.65	72.71	67.17	48.97	50.70	30.77	21.03
(6) L_{CCL} + L_{KL}	39.47	40.75	22.65	90.00	81.03	78.20	85.64	78.30	75.42	48.52	32.15	25.60
(7) DMvCGSA	**41.82**	**42.53**	**24.35**	**90.48**	**82.44**	**79.59**	**98.60**	**95.14**	**96.54**	**52.79**	**35.40**	**27.36**
Methods	Caltech-2V			Wiki			HandWritten			VOC		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
(1) L_{SGCL}	47.57	40.70	24.47	46.71	44.69	16.42	70.03	70.94	45.92	27.79	31.19	6.49
(2) L_{CCL}	57.05	41.77	34.22	51.39	44.29	33.66	83.18	83.86	76.65	39.81	45.26	28.04
(3) L_{KL}	37.33	27.37	19.04	47.15	44.51	30.72	60.13	66.76	51.25	31.61	36.85	22.27
(4) L_{SGCL} + L_{CCL}	58.83	48.99	39.44	54.73	48.20	37.12	82.08	81.10	73.90	44.55	54.07	34.15
(5) L_{SGCL} + L_{KL}	47.86	44.93	32.18	48.96	47.07	21.69	71.67	72.45	49.75	36.56	39.11	11.76
(6) L_{CCL} + L_{KL}	55.95	41.67	34.25	56.20	51.32	41.22	82.33	82.42	74.91	47.36	55.08	36.99
(7) DMvCGSA	**63.52**	**52.28**	**43.47**	**61.13**	**54.40**	**45.06**	**85.08**	**84.44**	**78.32**	**57.04**	**59.61**	**50.97**

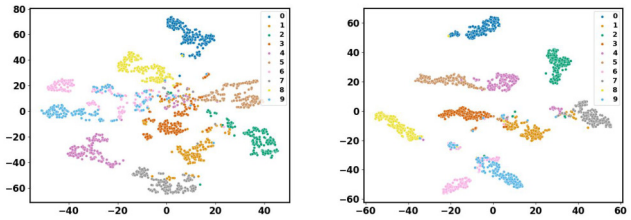
method usually achieves higher clustering performance than most state-of-the-art methods, which is consistent with most previous observations in two-view clustering and demonstrates the promising generality and scalability of the proposed method for multi-view clustering. It is also noted that our proposed DMvCGSA obtains slightly lower metric values than does CVCL and SEM on the Fashion dataset. This is possibly because CVCL and SEM employs a two-stage training strategy, including pretraining and fine-tuning, which can better capture relatively simple structural features with limited clusters for the Fashion samples by pretraining and further improve the clustering performance via fine-tuning, whereas our proposed DMvCGSA is an end-to-end scheme that cannot achieve a feature learning performance comparable to that of the two-stage-based CVCL and SEM method.

D. Ablation Studies

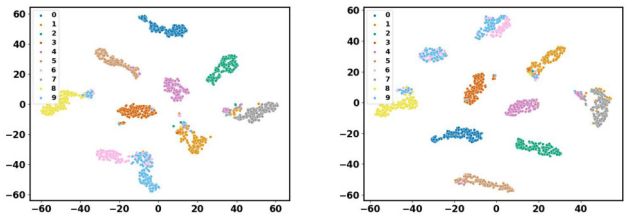
Our proposed method mainly consists of the similarity-guided instance-level contrastive learning, the cluster-level contrastive learning, and the multiple self-supervised learning components, corresponding to the losses of the L_{SGCL}, L_{CCL}, and L_{KL}, respectively. To evaluate the effectiveness of different components, we retain one and two of the three losses in the objective learning function and compare them with the proposed method by combining all three losses. Table IV tabulates the clustering results obtained by different variants of the proposed method. The variants that combine two losses usually perform better than those that include only one loss, and the proposed method achieves the best performance when the three losses are all considered during clustering, demonstrating the effectiveness of each component of the proposed method. The clustering performance can also be significantly improved when the instance-level and cluster-level contrastive learning modules are embedded. This is because the instance-level and cluster-level contrastive learning modules of the proposed method can effectively exploit the consistency information from multi-view data, avoiding errors caused by redundant information such that the robustness and reliability of the clustering results are enhanced.

Moreover, to test the GCN-AutoEncoder of the proposed method, we replace GCN-AutoEncoder with the autoencoder

Authorized licensed use limited to: NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS. Downloaded on August 20,2026 at 14:42:30 UTC from IEEE Xplore. Restrictions apply.



(a) The 1-th iteration (b) The 50-th iteration



(c) The 150-th iteration (d) The 500-th iteration

Fig. 2. t-SNE visualization of the latent representations obtained by the proposed DMvCGSA method with (a) epoch =0, (b) epoch =50, (c) epoch =150, and (d) epoch =500 on the HandWritten database.

TABLE V
ABLATION STUDY OF THE GCN-AUTOENCODER MODULE

Dataset	Model	ACC	NMI	ARI
Scene-15	(w/o) AE	38.73	41.24	24.29
	(w/o) GCN	39.89	42.00	24.09
	GCN-AutoEncoder	**41.82**	**42.53**	**24.35**
MSRC-V1	(w/o) AE	85.78	76.23	71.75
	(w/o) GCN	73.65	65.61	56.02
	GCN-AutoEncoder	**90.48**	**82.44**	**79.59**
BDGP	(w/o) AE	98.49	94.92	96.29
	(w/o) GCN	96.76	92.42	92.30
	GCN-AutoEncoder	**98.60**	**95.14**	**96.54**
Reuters	(w/o) AE	51.07	28.51	24.34
	(w/o) GCN	49.52	31.38	27.14
	GCN-AutoEncoder	**52.79**	**35.40**	**27.36**
Caltech-2V	(w/o) AE	22.74	7.26	3.56
	(w/o) GCN	54.40	40.29	32.87
	GCN-AutoEncoder	**63.52**	**52.28**	**43.47**
Wiki	(w/o) AE	54.66	46.02	36.88
	(w/o) GCN	56.01	54.08	43.28
	GCN-AutoEncoder	**61.13**	**54.40**	**45.06**
HandWritten	(w/o) AE	82.62	82.45	75.92
	(w/o) GCN	82.38	82.53	74.62
	GCN-AutoEncoder	**85.08**	**84.44**	**78.32**
VOC	(w/o) AE	48.09	50.04	38.58
	(w/o) GCN	47.68	57.37	36.71
	GCN-AutoEncoder	**57.04**	**59.61**	**50.97**

alone (referred to as (w/o) GCN) and the GCN alone (referred to as (w/o) AE), respectively, and tabulate the comparative multi-view clustering results in Table V. As can be seen, the proposed method embedded with GCN-AutoEncoder consistently outperforms the other two variants on all datasets. This is because GCN-AutoEncoder can simultaneously capture the intrinsic attribute features via the autoencoder and the high-order structural information via the GCN such that more informative representations can be exploited for multi-view clustering. It is noted that there is a significant performance drop on the Caltech-2V dataset when only the GCN layers are used. This is possibly because the Caltech-2V samples have complex scenarios and diverse feature distributions, making the clustering accuracy heavily relying on the attribute information of multi-view samples extracted by the autoencoder.

E. Parameter Sensitivity

Our proposed method contains six hyperparameters, including the number of neighbors (i.e., K), the threshold for similarity-guided instance-level contrastive learning (i.e., t), the balance coefficient (i.e., μ), and the trade-off parameters (i.e., λ_1, λ_2, and λ_3). To the best of our knowledge, there is no effective way find the optimal values of these parameters directly for different kinds of datasets due to the diverse characteristics of the samples. In this study, we mainly analyze the performance of the proposed method versus different values for each parameter on the MSRC-V1 dataset, as shown in Fig. 3.

Specifically, Fig. 3(a) shows the clustering performance plotted versus different values of K, and the proposed method usually achieves the best performance when K is set to approximately 8-11. This is because a small value of K usually results in a sparse adjacency matrix, discarding the affinity information among genuine neighboring samples, while a large K value possibly produces inaccurate neighborhood relationships among samples.

Fig. 3(b) depicts the clustering performance of the proposed method plotted versus different values of the threshold t selected from the candidate set of $\{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, and the best clustering performance is achieved when t is 0.8. This is because a small value of t possibly discards the true negative pairs. As a result, the diversity of the negative sample set is significantly reduced, making the proposed method hard to obtain clear boundaries of different clusters. Conversely, a large value of the t possibly misclassifies the intra-cluster samples as negative pairs, resulting in the destruction of the intra-cluster structure.

Moreover, to study the impact of the balance coefficient μ, Fig. 3(c) shows the clustering performance of the proposed method with different values of μ selected from $\{0.0, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$, in which $\mu = 0.0$ indicates that only the GCN is used and $\mu = 1.0$ indicates that only the autoencoder is used during feature encoding. Our proposed method usually achieves better clustering performance when μ is 0.5, demonstrating the effectiveness of the developed GCN-Autoencoder module for exploring both the structural and attribute information in multi-view feature representations. Theoretically, our proposed method with a small value of the μ focuses most on learning graph structure information. This makes the node embeddings of the multi-view data similar, possibly resulting in over-smoothing problem. By contrast, the proposed method with a large value of the μ focuses on learning the attribute information of multi-view data. As a result, it is hard to obtain the high-order structural information, thereby limiting the clustering performance.

Authorized licensed use limited to: NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS. Downloaded on August 20,2026 at 14:42:30 UTC from IEEE Xplore. Restrictions apply.

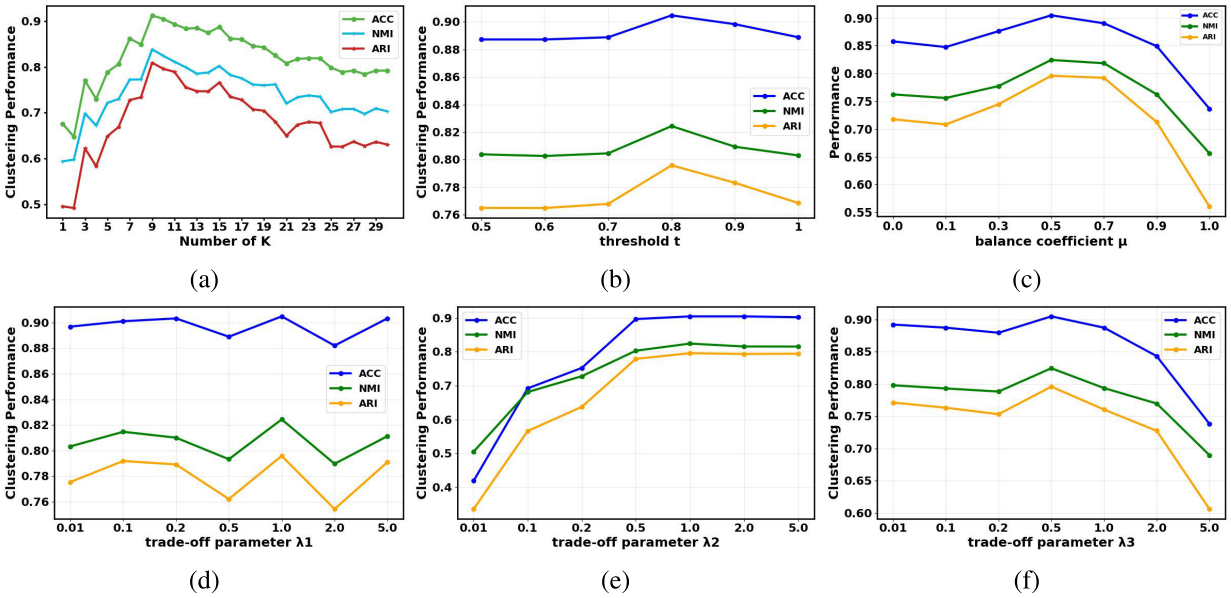


Fig. 3. Clustering performance of the proposed method in terms of the ACC, NMI and ARI versus different values: (a) number of neighbors K, (b) threshold t, (c) balance coefficient μ and (d) - (f) trade-off parameters λ_1, λ_2, and λ_3.

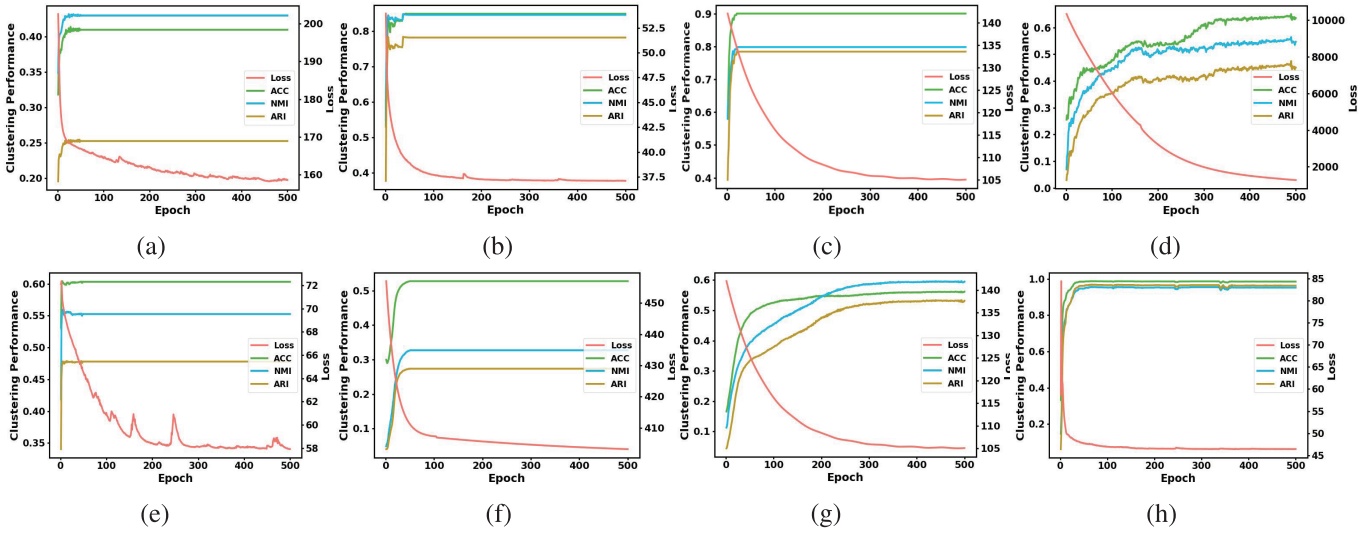


Fig. 4. Clustering results and corresponding objective function values of the proposed method on the (a) Scene-15, (b) HandWritten, (c) MSRC-V1, (d) Caltech-2V, (e) Wiki, (f) Reuters, (g) VOC and (h) BDGP datasets.

Finally, Figs. 3(d), (e), and (f) present the clustering results of the proposed method with different values of λ_1, λ_2 and λ_3, respectively. The clustering results of the proposed method fluctuate slightly with different values of λ_1, showing less sensitivity to the parameter, whereas the proposed method usually achieves the best performance when a relatively large λ_2 value is set. The main reason is that cluster-level contrastive learning module can directly capture the global cluster structure information, thereby learning more clustering-friendly feature representations. In addition, the proposed method achieves a better performance when λ_3 is set to a smaller value, because the multiple self-supervised loss with such a small coefficient value can better align the multi-view clustering results with a high-confidence target distribution.

F. Convergence Analysis

Fig. 4 shows the multi-view clustering results and the corresponding objective function values on all datasets. It is observed that the proposed method can usually converge in about 200 to 400 epochs, demonstrating the rapid convergence speed of the proposed method.

V. Conclusion

In this paper, we propose a new deep multi-view clustering network via graph structure awareness, named DMvCGSA, which first extracts view-specific structural information and attribute features via the integration of GCN with autoencoder, and then performs double contrastive learning at both the instance and cluster levels to make the multi-view

Authorized licensed use limited to: NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS. Downloaded on August 20,2026 at 14:42:30 UTC from IEEE Xplore. Restrictions apply.

representation collaborative and view-specific clustering results consistent, resulting in improved multi-view clustering performance. Extensive experimental results on twelve widely used datasets clearly demonstrate the effectiveness and scalability of the proposed method for multi-view clustering. For future studies, extending our proposed method to an incomplete multi-view clustering scenario may be an interesting direction.

REFERENCES

[1] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[2] Z. Ding, M. Shao, and Y. Fu, "Robust multi-view representation: A unified perspective from multi-view learning to domain adaption," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 5434–5440.

[3] Q. Wang, Z. Tao, Q. Gao, and L. Jiao, "Multi-view subspace clustering via structured multi-pathway network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 5, pp. 7244–7250, May 2024.

[4] Z. Lin, Z. Kang, L. Zhang, and L. Tian, "Multi-view attributed graph clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 2, pp. 1872–1880, Feb. 2023.

[5] G. Chao, S. Sun, and J. Bi, "A survey on multiview clustering," *IEEE Trans. Artif. Intell.*, vol. 2, no. 2, pp. 146–168, Apr. 2021.

[6] H. Zhao, Z. Ding, and Y. Fu, "Multi-view clustering via deep matrix factorization," in *Proc. 31st Conf. Artif. Intell. (AAAI)*, Feb. 2017, pp. 2921–2927.

[7] Y. Wang, X. Lin, L. Wu, W. Zhang, Q. Zhang, and X. Huang, "Robust subspace clustering for multi-view data by exploiting correlation consensus," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3939–3949, Nov. 2015.

[8] F. Nie, J. Li, and X. Li, "Self-weighted multiview clustering with multiple graphs," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 2564–2570.

[9] W. Zhao, C. Xu, Z. Guan, and Y. Liu, "Multiview concept learning via deep matrix factorization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 814–825, Feb. 2021.

[10] J. Cui, Y. Li, H. Huang, and J. Wen, "Dual contrast-driven deep multi-view clustering," *IEEE Trans. Image Process.*, vol. 33, pp. 4753–4764, 2024.

[11] Q. Wang, J. Cheng, Q. Gao, G. Zhao, and L. Jiao, "Deep multi-view subspace clustering with unified and discriminative learning," *IEEE Trans. Multimedia*, vol. 23, pp. 3483–3493, 2021.

[12] Z. Ding and Y. Fu, "Robust multi-view subspace learning through dual low-rank decompositions," in *Proc. 13th AAAI Conf. Artif. Intell.*, vol. 30, Feb. 2016, pp. 1181–1187.

[13] Y. Ren et al., "Deep clustering: A comprehensive survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 4, pp. 5858–5878, Apr. 2025.

[14] J. Wen, Z. Zhang, Y. Xu, B. Zhang, L. Fei, and G.-S. Xie, "CDIMC-Net: Cognitive deep incomplete multi-view clustering network," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 3230–3236.

[15] Q. Wang, Z. Ding, Z. Tao, Q. Gao, and Y. Fu, "Generative partial multi-view clustering with adaptive fusion and cycle consistency," *IEEE Trans. Image Process.*, vol. 30, pp. 1771–1783, 2021.

[16] S. Wang, C. Li, Y. Li, Y. Yuan, and G. Wang, "Self-supervised information bottleneck for deep multi-view subspace clustering," *IEEE Trans. Image Process.*, vol. 32, pp. 1555–1567, 2023.

[17] Q. Gao, H. Lian, Q. Wang, and G. Sun, "Cross-modal subspace clustering via deep canonical correlation analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 3938–3945.

[18] S. Luo, C. Zhang, W. Zhang, and X. Cao, "Consistent and specific multi-view subspace clustering," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2018, vol. 32, no. 1, pp. 3731–3737.

[19] J. Cheng, Q. Wang, Z. Tao, D. Xie, and Q. Gao, "Multi-view attribute graph convolution networks for clustering," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 2973–2979.

[20] 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.*, Jun. 2022, pp. 16051–16060.

[21] J. Wang, S. Feng, G. Lyu, and Z. Gu, "Triple-granularity contrastive learning for deep multi-view subspace clustering," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 2994–3002.

[22] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–14.

[23] E. Pan and K. Zhao, "Beyond homophily: Reconstructing structure for graph-agnostic clustering," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2023, pp. 1–10.

[24] T. N. Kipf and M. Welling, "Variational graph auto-encoders," 2016, *arXiv:1611.07308*.

[25] X. Li, H. Zhang, and R. Zhang, "Adaptive graph auto-encoder for general data clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9725–9732, Dec. 2022.

[26] J. Wang, S. Feng, G. Lyu, and J. Yuan, "SURER: Structure-adaptive unified graph neural network for multi-view clustering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, Mar. 2024, pp. 15520–15527.

[27] Y. Wang, D. Chang, Z. Fu, J. Wen, and Y. Zhao, "Incomplete multiview clustering via cross-view relation transfer," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 367–378, Jan. 2023.

[28] 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.

[29] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 10, pp. 8547–8555.

[30] S. Hu, C. Zhang, G. Zou, Z. Lou, and Y. Ye, "Deep multiview clustering by pseudo-label guided contrastive learning and dual correlation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 2, pp. 3646–3658, Feb. 2025.

[31] D. Bo, X. Wang, C. Shi, M. Zhu, E. Lu, and P. Cui, "Structural deep clustering network," in *Proc. Web Conf.*, Apr. 2020, pp. 1400–1410.

[32] W. Hu, T. Miyato, S. Tokui, E. Matsumoto, and M. Sugiyama, "Learning discrete representations via information maximizing self-augmented training," in *Proc. 34th Int. Conf. Mach. Learn.*, Feb. 2017, pp. 1158–1167.

[33] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 478–487.

[34] F. Li and P. Pietro, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 524–531.

[35] G. Chao, Y. Jiang, and D. Chu, "Incomplete contrastive multi-view clustering with high-confidence guiding," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 10, pp. 11221–11229.

[36] J. Shotton, A. Blake, and R. Cipolla, "Multiscale categorical object recognition using contour fragments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1270–1281, Jul. 2008.

[37] 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, Jun. 2012.

[38] M. Amini, N. Usunier, and C. Goutte, "Learning from multiple partially observed views—An application to multilingual text categorization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Dec. 2009, pp. 28–36.

[39] M. Yang, Y. Li, Z. Huang, Z. Liu, P. Hu, and X. Peng, "Partially view-aligned representation learning with noise-robust contrastive loss," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 1134–1143.

[40] 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. Conf. Comput. Vis. Pattern Recognit. Workshop*, 2004, p. 178.

[41] C. Yang, Z. Liu, D. Zhao, M. Sun, and E. Chang, "Network representation learning with rich text information," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 2111–2117.

[42] S. Perkins and J. Theiler, "Online feature selection using grafting," in *Proc. 20th Int. Conf. Mach. Learn.*, 2003, pp. 592–599.

[43] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009.

[44] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," 2017, *arXiv:1708.07747*.

[45] J. Xu et al., "Multi-VAE: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9214–9223.

[46] M. Brbi, M. Pikorec, V. Vidulin, A. Kriko, T. Muc, and F. Supek, "The landscape of microbial phenotypic traits and associated genes," *Nucleic acids Res.*, vol. 44, no. 21, pp. 10074–10090, 2016.

Authorized licensed use limited to: NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS. Downloaded on August 20,2026 at 14:42:30 UTC from IEEE Xplore. Restrictions apply.

[47] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Jan. 1988.
[48] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, May 1998.
[49] J. A. Hartigan and M. A. Wong, "A K-means clustering algorithm," *Appl. Statist.*, vol. 28, no. 1, pp. 100–108, 1979.
[50] 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, Sep. 2021.
[51] D. J. Trosten, S. Løkse, R. Jenssen, and M. Kampffmeyer, "Reconsidering representation alignment for multi-view clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1255–1265.
[52] X. Yang et al., "DealMVC: Dual contrastive calibration for multi-view clustering," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 337–346.
[53] J. Chen, H. Mao, W. L. Woo, and X. Peng, "Deep multiview clustering by contrasting cluster assignments," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 16706–16715.
[54] J. Xu et al., "Self-weighted contrastive learning among multiple views for mitigating representation degeneration," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 36, 2023, pp. 1–10.
[55] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

Lunke Fei (Senior Member, IEEE) received the Ph.D. degree in computer science and technology from Harbin Institute of Technology (Shenzhen), Shenzhen, China, in 2016. He was a Postdoctoral Fellow with the Department of Computer and Information Science, University of Macau, Macau, China. He is currently an Associate Professor with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China. His research interests include pattern recognition, biometrics, computer vision, and machine learning.

Junlin He received the B.S. degree from the School of Computer Science, Hunan University of Technology and Business, Changsha, China, in 2023. He is currently pursuing the M.S. degree with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China. His main research interests include multi-view clustering and machine learning.

Qi Zhu received the B.S., M.S., and Ph.D. degrees from Harbin Institute of Technology in 2007, 2010, and 2014, respectively. He is currently a Professor with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. He has published over 100 scientific papers in refereed international journal and conference proceedings. His research interests include pattern recognition, medical image analysis, and biometrics.

Shuping Zhao (Member, IEEE) received the M.S. degree in information science from South China Normal University, Guangzhou, China, in 2016, and the Ph.D. degree in computer science from the Department of Computer and Information Science, University of Macau, Taipa, Macau, in 2020. He is currently with the School of Computer Science, Guangdong University of Technology, Guangzhou. He has published a number of technical papers at prestigious international journals, such as IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, and international conferences. His research interests include biometrics, machine learning, pattern recognition, and image processing.

Jie Wen (Senior Member, IEEE) received the Ph.D. degree in computer science and technology from Harbin Institute of Technology (Shenzhen) in 2019. He is currently an Associate Professor with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen). He has authored or co-authored more than 100 technical papers at prestigious international journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, NeurIPS, ICML, CVPR, AAAI, IJCAI, and ACM MM. His research interests include image and video enhancement, pattern recognition, and machine learning. One conference paper received the distinguished Paper Award from AAAI 2023. He serves/served as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, *Pattern Recognition*, and *International Journal of Image and Graphics*, an Area Editor for *Information Fusion*, and an Area Chair for NeurIPS, ICML, and ACM MM. He featured among the "World's Top 2% Scientists List" in 2021 and 2024. For more information, please refer to homepage: https://sites.google.com/view/jerry-wen-hit/home

Yong Xu (Senior Member, IEEE) received the B.S. and M.S. degrees in 1994 and 1997, respectively, and the Ph.D. degree in pattern recognition and intelligence system from Nanjing University of Science and Technology, Nanjing, China, in 2005. He is currently with the Bio-Computing Research Center, Harbin Institute of Technology (Shenzhen), and Shenzhen Key Laboratory of Visual Object Detection and Recognition, Shenzhen, China. His current research interests include pattern recognition, biometrics, bioinformatics, machine learning, image processing, and video analysis.

Authorized licensed use limited to: NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS. Downloaded on August 20,2026 at 14:42:30 UTC from IEEE Xplore. Restrictions apply.