Self-Weighted Contrastive Fusion for Deep Multi-View Clustering

Song Wu, Yan Zheng, Yazhou Ren, *Member, IEEE*, Jing He, Xiaorong Pu, Shudong Huang, Zhifeng Hao, *Senior Member, IEEE*, and Lifang He, *Member, IEEE*

***Abstract*—Multi-view clustering can explore consensus information from multiple views and has attracted increasing attention in the past two decades. However, existing works face two major challenges: i) how to deal with the conflict between learning view-consensus information and reconstructing inconsistent view-private information and ii) how to mitigate representation degeneration caused by implementing the consistency objective for multi-view data. To address these challenges, we propose a novel framework of self-weighted contrastive fusion for deep multi-view clustering (SCMVC). First, our method establishes a hierarchical feature fusion framework, effectively segregating the consistency objective from the reconstruction objective. Then, multi-view contrastive fusion is implemented via maximizing consistency expression between the view-consensus representation and global representation, fully exploring the view consistency and complementary. More importantly, we propose to measure the discrepancy between pairwise representations, and then introduce a self-weighting method, which adaptively strengthens useful views in feature fusion and weakens unreliable views, to mitigate representation degeneration. Extensive experiments on nine public datasets demonstrate that our proposed method achieves state-of-the-art clustering performance.**

***Index Terms*—Multi-view contrastive learning, multi-view clustering, deep clustering, representation degeneration.**

I. Introduction

With the rapid development of multimedia applications, a large amount of data is being collected from various sources or described with diverse attributes. In particular, these data generally lack label information. For instance, in the case of a video, it may include images captured from different cameras, audio with varying voices, and text descriptions. To explore useful consistent and complementary information among multiple views in an unsupervised way, multi-view clustering (MVC) [1], [2], [3], [4] aims to integrate data from different sources to gain a more comprehensive understanding of the underlying phenomena.

In the field of multi-view clustering (MVC), deep multi-view clustering methods [5], [6] have been proven to achieve superior clustering performance owing to the powerful representation learning ability of deep networks. Specifically, these methods [7], [8] employ a view-specific encoder network to learn the salient features for each view. Then, these learned view representations are further fused to obtain a more discriminative global feature that can be divided into different categories based on the complementary information across all views. Despite considerable progress has been made in recent years within the realm of deep multi-view clustering, there remain two major challenges: (i) *how to deal with the conflict between learning common view-consensus information and reconstructing inconsistent view-private information*, and (ii) *how to mitigate representation degeneration caused by implementing the consistency objective for mutli-view data.*

More specifically, multi-view data generally contains two types of information, *i.e.*, *the consensus information* across all views, and *the inconsistent view-private information* about the individual view. In MVC, an intuitive idea is to capture as much consensus information as possible across all views, thereby exploring more discriminating clustering structures [9]. In light of this, most deep MVC methods, *e.g.*, [8], [10], conduct the consistency objective on the latent features to unveil view consistency. However, they tend to ignore that the reconstruction objective retained in the same feature space might compel the salient features to redundantly reconstruct meaningless private information. To elaborate, the former tries to learn the consensus features across all views as much as possible, while the latter wants to maintain the invariance between inputs and outputs for

Manuscript received 12 December 2023; revised 19 March 2024; accepted 26 March 2024. Date of publication 16 April 2024; date of current version 26 August 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFC2004300 and Grant 2020YFC2004302, in part by Shenzhen Science and Technology Program under Grant JCYJ20230807120010021 and Grant JCYJ20230807115959041, in part by National Science Foundation under Grant 61971052, and in part by the open project of Sichuan Provincial Key Laboratory of Philosophy and Social Science for Language Intelligence in Special Education under Grant YYZN-2023-3. The work of Lifang He was supported by NSF under Grant MRI-2215789, Grant IIS-1909879, and Grant IIS-2319451, in part by NIH under Grant R21EY034179, and in part by Lehigh's grants under Accelerator and CORE. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. B. Bao. *(Corresponding author: Yazhou Ren.)*

Song Wu and Yan Zheng are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: songwu.work@outlook.com; yan9zheng9@gmail.com).

Yazhou Ren and Xiaorong Pu are with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518000, China (e-mail: yazhou.ren@uestc.edu.cn; puxiaor@uestc.edu.cn).

Jing He is with the Nuffield Department of Clinical Neurosciences, University of Oxford, OX3 9DU Oxford, U.K. (e-mail: lotusjing@gmail.com).

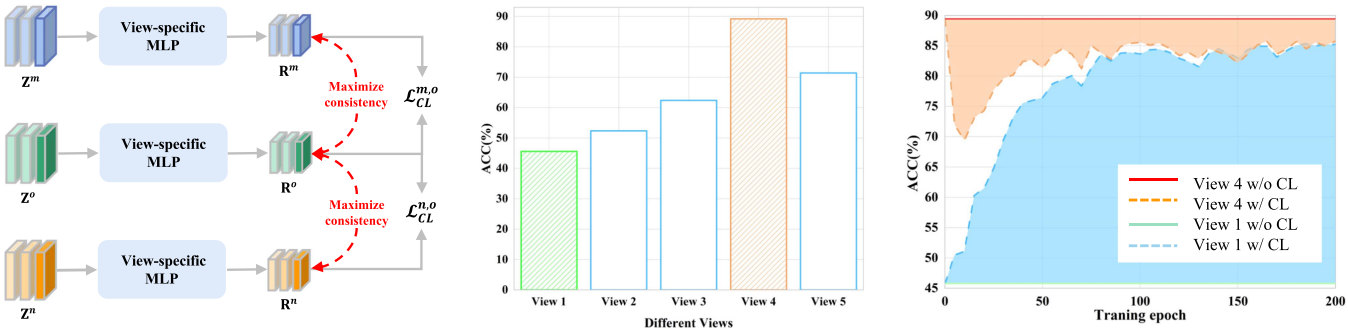
Shudong Huang is with the College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: huangsd@scu.edu.cn).

Zhifeng Hao is with the College of Science, Shantou University, Shantou 515063, China (e-mail: haozhifeng@stu.edu.cn).

Lifang He is with the Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA 18015 USA (e-mail: lih319@lehigh.edu).

The code is available at https://github.com/SongwuJob/SCMVC.

Digital Object Identifier 10.1109/TMM.2024.3387298



(a) A typical **CL** framework (b) The quality of different views (c) Representation degradation

Fig. 1. (a) Typical MCL framework: the latent embeddings $\{\mathbf{Z}^v\}_{v=1}^M$ are projected into the consistent feature spaces $\{\mathbf{R}^v\}_{v=1}^M$, where the contrastive learning is implemented among different views, *i.e.*, $\mathcal{L}_{CL}^{m,o}$ and $\mathcal{L}_{CL}^{n,o}$. (b) Clustering accuracy of individual views on Caltech-5 V dataset. (c) Clustering accuracy of view 1 and view 4 with typical MCL framework. The high-quality views will be forced to align with the low-quality views.

the individual view. This inconsistent conflict severely limits MVC methods.

To handle the above challenges, contrastive multi-view clustering methods [9], [11] have been proposed, emphasizing the alignment of representations from each view to mine consensus information. Despite achieving satisfactory results, we found that the excessive pursuit of view consistency might cause *representation degeneration* that the high-quality views would be forced to align with the low-quality views to achieve maximum view consistency. This side effect restricts the effectiveness of multi-view clustering tasks (see Fig. 1). Moreover, the global complementary information is often discarded during contrastive learning [12]. The lack of complementary semantics might further exacerbate representation degeneration, thereby leading to the inability to capture sufficient discriminative information.

To address the above issues, we propose a novel framework of self-weighted contrastive fusion for deep multi-view clustering (SCMVC). Specifically, focusing on the challenge (i), we establish a hierarchical feature fusion framework to avoid the reconstruction loss directly acting on consensus feature learning. First, we leverage the autoencoders to learn the low-level features from raw data. Subsequently, two MLPs are stacked to separate the consistent feature learning from the reconstruction objective, where a linear MLP is used to mine view-consensus information for each view, while another nonlinear MLP conducts feature fusion on all latent embeddings to fully explore the complementary information. Motivated by the insight that the salient representations of the same sample from different views are typically similar, we conduct multi-view contrastive fusion between view-consensus features and global features to achieve the consistency objective. Considering the challenge (ii), we propose to first measure the discrepancy among pairwise representations, and then adaptively strengthen useful views in feature fusion, weakening unreliable views. In this way, high-quality views with informative semantics would dominate the feature fusion, while significantly reducing the impact of low-quality views. In summary, our key contributions are as follows:

- We propose a hierarchical feature fusion framework where different objectives are conducted in different feature spaces. In this way, our method can effectively explore the consensus information for each view, and further learn the global discriminative representation for the downstream task.
- We propose a novel self-weighted multi-view contrastive fusion paradigm, which can adaptively strengthen useful views with informative semantics in feature fusion while reducing the impact of unreliable views.
- Extensive experiments are conducted on nine public datasets, and the results demonstrate the state-of-the-art clustering performance of our proposed method.

II. Related Work

A. Multi-View Clustering

In this paper, we roughly divide the existing MVC methods into four categories: (1) Subspace-based multi-view clustering [13], [14]. In [15], latent subspace representations, which are more accurate and robust, are learned by leveraging the complementarity of multiple views. Liu et al. [16] combine anchor learning and graph construction into a uniform framework. Particularly, the algorithm directly outputs the clustering via graph connectivity constraint. (2) Matrix factorization based multi-view clustering [17]. Non-negative matrix factorization is used to decompose each view into low-rank matrices, and then the data are clustered in a low-dimensional space [18]. Wei et al. [19] propose a deep matrix factorization based solution, where multi-view data matrices are factorized into multiple representational subspaces layer by layer. (3) Graph-based multi-view clustering [20], [21]. Many MVC methods aim to generate more significant clustering representations by introducing topological information [22]. In [23], the graph autoencoder is used to learn latent clustering representations, where one informative graph view is employed, and latent representations are reconstructed into multiple graph views. (4) Deep embedded multi-view clustering [24]. One of the most representative works is deep embedded clustering DEC [25], which jointly learns the clustering assignments and embedded features of autoencoders. Based on this, improved DEC [26] introduces a trade-off between clustering and reconstruction objectives to prevent the

collapse of deep models. Furthermore, Yan et al. [27] further introduce the transformer architecture for deep multi-view clustering task, where the structure relationship of all samples is fully explored.

B. Multi-View Contrastive Learning

Contrastive learning is a novel unsupervised representation learning method that aims to learn feature representations by comparing the similarity or difference between different data points [28], [29]. In computer vision, the contrastive learning paradigm has been widely used due to its effective feature learning ability [30], [31]. For instance, Zhong et al. [32] lift the traditional instance-level consistency to the cluster-level consistency via contrastive learning. Particularly, multi-view contrastive learning (MCL) aims to handle multi-view data widely existed in multimedia applications, attracting increasing attention [33], [34]. Ke et al. [12] conduct contrastive fusion from multiple views, and the characteristics of view-specific representations are maintained. Xu et al. [9] explore how to learn the view-consensus representation and avoid the impact of the view-private information, where different levels of features are learned via contrastive learning. In [35], a dual mutual information constrained clustering method is proposed, where the mutual information across all the dimensionalities is minimized, and that of the similar instance pairs is maximized. Although superior results have been achieved in many cases, we found that most previous works often ignore the representation degeneration problem that the high-quality views would be forced to align with the low-quality views, like Fig. 1. Focusing on addressing this issue, in this paper, we propose a self-weighted contrastive fusion framework (SCMVC).

III. The Proposed Method

Problem Statement: Given a multi-view dataset $\{\mathbf{X}^v\}_{v=1}^M$ with N samples across M views, where $\mathbf{X}^v = \{\mathbf{X}_1^v; \mathbf{X}_2^v; \ldots; \mathbf{X}_N^v\} \in \mathbb{R}^{N\times D_v}$, D_v denotes the dimension of the raw features in the v-th view. Multi-view clustering aims to partition N instances into k clusters. In order to enhance clarity and conciseness, the main symbols used in our study are listed in Table I.

TABLE I
Descriptions of Main Symbols Used in Our Study

Symbols	Descriptions
M	The number of views.
N	The number of samples.
$\mathbf{X}^v$	The raw feature in the v-th view.
$\mathbf{Z}^v$	The low-level feature in the v-th view.
$\mathbf{R}^v$	The view-consensus feature in the v-th view.
$\mathbf{H}$	The global feature for clustering.
$\mathcal{W}^v$	The adaptive weight for the view v.
D_v/d_v	The dimension of $\mathbf{X}^v$ and $\mathbf{Z}^v$.
d_r/d_h	The dimension of $\mathbf{R}^v$ and $\mathbf{H}$.

A. Motivation

In general, multi-view datasets are susceptible to containing noise and redundant information. Hence, the mainstream methods generally implement self-supervised autoencoder models, *e.g.*, AE [36], VAE [37], and MAE [38], to learn salient representations from raw features. Specifically, for the v-th view, let $E^v(\mathbf{X}^v;\theta^v)$ and $D^v(\mathbf{Z}^v;\phi^v)$ represent multi-layer nonlinear encoder and decoder, where θ^v and ϕ^v are the learnable parameters of autoencoder networks, denote $\mathbf{Z}^v = E^v(\mathbf{X}^v) \in \mathbb{R}^{N\times d_v}$ as the latent embedding in d_v-dimensional feature space. Then, the autoencoders are optimized via forcing the decoded output $\hat{\mathbf{X}}^v = D^v(\mathbf{Z}^v) \in \mathbb{R}^{N\times D_v}$ to be consistent with the original input $\mathbf{X}^v$, so the reconstruction objective $\mathcal{L}_{\boldsymbol{Z}}$ can be formulated as:

$$\mathcal{L}_{\boldsymbol{Z}} = \sum_{v=1}^{M} \mathcal{L}_{\boldsymbol{Z}}^v = \sum_{v=1}^{M} \left\| \mathbf{X}^v - D^v\left(E^v\left(\mathbf{X}^v\right)\right)\right\|_F^2. \tag{1}$$

Despite the popularity of autoencoder models, its effectiveness actually is constrained by two major factors: (i) **Impact of view-private information:** In (1), the $\mathcal{L}_{\boldsymbol{Z}}$ aims to reconstruct the latent embeddings $\mathbf{Z}^v$ consistent with the input, which would introduce much view-private information. They are meaningless, and even lead to model collapse. (ii) **Lack of information interaction:** The autoencoder is limited to its own view information, where it lacks cross-view interaction, and complementary information across all views is ignored. To address the aforementioned constraints, multi-view contrastive learning (*e.g.*, CoMVC [11] and MFLVC [9]) aims to mine the consistent information for multiple views. Specifically, as shown in Fig. 1(a), we denote $\mathcal{R}(\mathbf{Z}^v;\Psi)$ as a feature MLP acted on $\{\mathbf{Z}^v\}_{v=1}^M$, to filter out meaningless private information for all views, and $\mathcal{L}_{CL}^{m,n}(\mathcal{R}(\mathbf{Z}^m),\mathcal{R}(\mathbf{Z}^n))$ represents a view contrastive loss. Then, the overall objective is conducted by minimizing the following loss function:

$$\sum_{v} \mathcal{L}_{\boldsymbol{Z}}\left(\mathbf{X}^v, \hat{\mathbf{X}}^v\right) + \lambda \sum_{m,n} \mathcal{L}_{\boldsymbol{CL}}^{m,n}\left(\mathcal{R}\left(\mathbf{Z}^m\right), \mathcal{R}\left(\mathbf{Z}^n\right)\right), \tag{2}$$

where the consistency objective is conducted by aligning latent feature spaces from different views, and $\lambda > 0$ denotes a trade-off coefficient. The view-consensus representations $\mathbf{R}^v = \mathcal{R}(\mathbf{Z}^v) \in \mathbb{R}^{N\times d_r}$ in d_r-dimensional feature space are used to the downstream task.

Nevertheless, as shown in Fig. 1(b)–(c), we found that MCL might lead to **representation degeneration** that the high-quality views would be forced to align with low-quality views. There are two main reasons: 1) Most previous works, like (2), conduct the consistency objective based on a prior condition that different views have semantic consistency. However, the characteristics and qualities inherent in different views typically exhibit significant variation. Low-quality views tend to limit the effectiveness of MCL. 2) The excessive pursuit of view consistency might

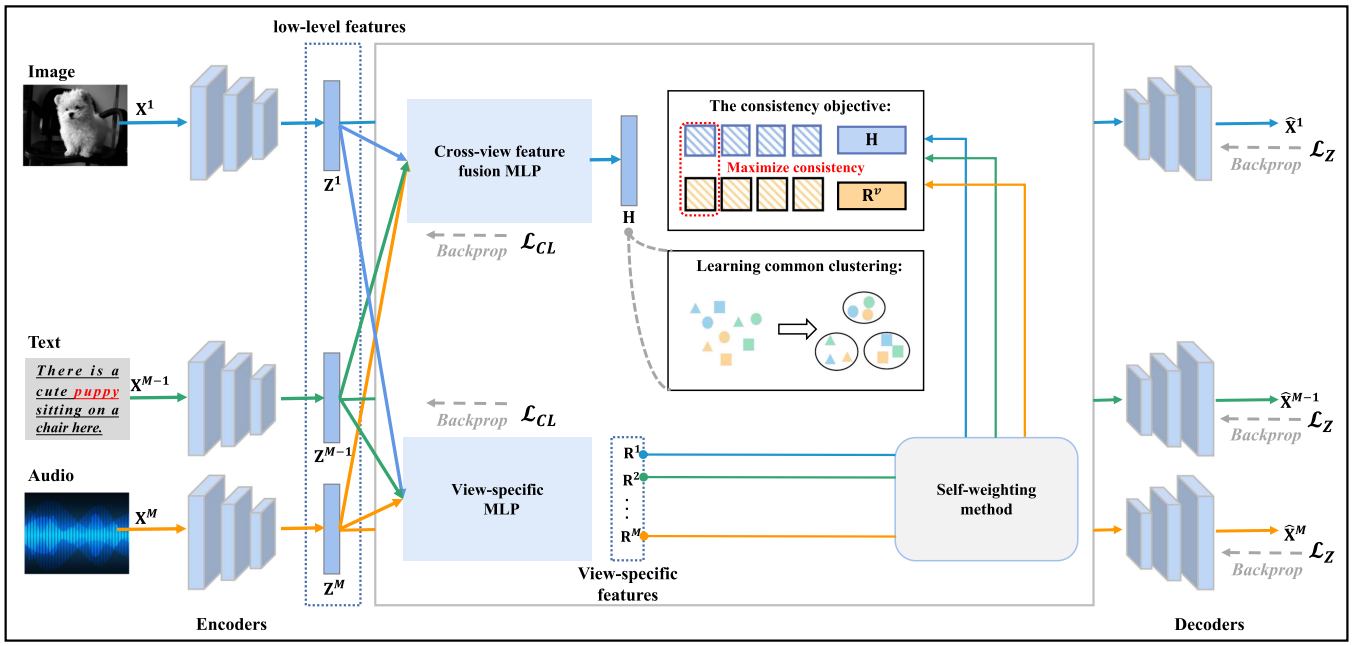


Fig. 2. Framework of SCMVC. We propose a hierarchical network architecture to separate the consistency objective from the reconstruction objective. Specifically, the feature learning autoencoders first project the raw data into a low-dimensional latent space $\{\mathbf{Z}^v\}_{v=1}^M$. Then, two feature MLPs learn view-consensus features $\{\mathbf{R}^v\}_{v=1}^M$ and global features $\mathbf{H}$, respectively. Particularly, a novel self-weighting method adaptively strengthens useful views in feature fusion, and weakens unreliable views, to implement multi-view contrastive fusion.

cause the model to discard complementary information, which would yield the final features that capture insufficient semantics.

To address these challenges, we propose a new framework of self-weighted contrastive fusion for deep multi-view clustering (SCMVC) as shown in Fig. 2. To fully explore cross-view complementary information, we extend the previous framework like Fig. 1(a), implementing global feature learning through the fusion of all latent features. Then, multi-view contrastive fusion is conducted by maximizing consistency expression between view-consensus features and global features. More importantly, to mitigate representation degeneration, we implement the consistency objective via a self-weighting method, which adaptively strengthens useful views, and reduces the impact of unreliable views. Overall, our optimization objective is:

$$\sum_v \mathcal{L}_Z\left(\mathbf{X}^v, \hat{\mathbf{X}}^v\right) + \sum_v \mathcal{W}^v \mathcal{L}_{CL}^v\left(\mathbf{R}^v, \mathbf{H}\right). \tag{3}$$

where $\mathcal{W}^v$ is an adaptive view weight. $\mathbf{H}$ and $\mathbf{R}^v$ denote the global representation and view-consensus representation, which will be introduced in the following section.

B. Self-Weighted Contrastive Fusion

As aforementioned, the features $\{\mathbf{Z}^v\}_{v=1}^M$ obtained by (1) mix both consensus and private information. To solve it, we propose to establish a hierarchical feature fusion framework. As depicted in Fig. 3, we first treat $\{\mathbf{Z}^v\}_{v=1}^M$ as low-level features, and stack a linear feature MLP $\mathcal{R}(\mathbf{Z}^v; \Psi)$ on $\{\mathbf{Z}^v\}_{v=1}^M$ to obtain view-consensus features $\{\mathbf{R}^v\}_{v=1}^M$, filtering out meaningless private information. Meanwhile, unlike the previous MCL works, like Fig. 1(a), which often ignore the complementary information,

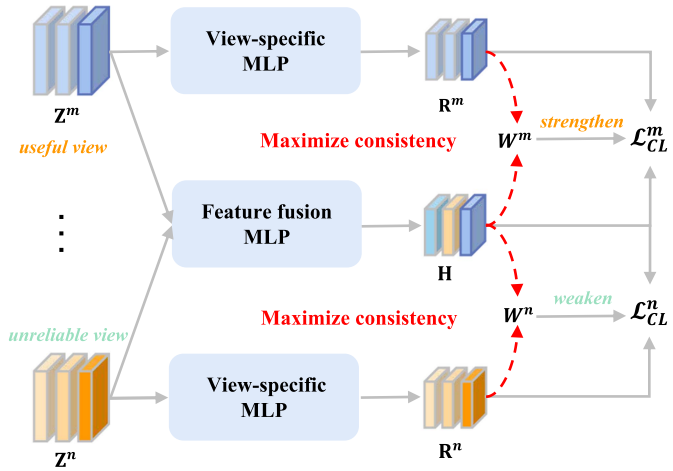


Fig. 3. Self-weighted contrastive fusion framework. $\mathbf{Z}$, $\mathbf{R}$, and $\mathbf{H}$ denote the low-level features, view-consensus features, and global features, respectively. The consistency objective (*i.e.*, $\mathcal{L}_{CL}^m$ and $\mathcal{L}_{CL}^n$) is implemented in a self-weighting manner.

we extend our approach to learn global features $\mathbf{H}$ by stacking another nonlinear fusion MLP on $\{\mathbf{Z}^v\}_{v=1}^M$. In this way, the gradients from the reconstruction objective in (1) cannot directly act on $\{\mathbf{R}^v\}_{v=1}^M$ and $\mathbf{H}$. The global representation $\mathbf{H}$ can be computed as:

$$\mathbf{H} = \mathcal{F}\left(\hat{\mathbf{Z}}; \Phi\right) = \mathcal{F}\left(\left[\mathbf{Z}^1, \mathbf{Z}^2, \ldots, \mathbf{Z}^M\right]; \Phi\right), \tag{4}$$

where $\hat{\mathbf{Z}} \in \mathbb{R}^{N \times d}$, $d = M \times d_v$, and $\mathbf{H} \in \mathbb{R}^{N \times d_h}$. We denote Ψ and Φ as the parameters of MLPs. To preserve the consistency between $\mathbf{H}$ and $\{\mathbf{R}^v\}_{v=1}^M$, we set $d_h = d_r$.

Inspired by MCL, we maximize the consistency expression between the features $\{\mathbf{R}^v\}_{v=1}^M$ and $\mathbf{H}$. The global features $\mathbf{H}$ can directly access consensus information from each view, rather than indirectly acquire common semantics through feature alignment. Then, the overall objective is:

$$\sum_v \mathcal{L}_{\boldsymbol{Z}}\left(\mathbf{X}^v, \hat{\mathbf{X}}^v\right) + \lambda \sum_v \mathcal{L}_{\boldsymbol{CL}}^v\left(\mathbf{R}^v, \mathbf{H}\right). \tag{5}$$

In consensus feature space, the learned global features $\mathbf{H}$ summarize the consensus information of each view, where these view-consensus representations $\{\mathbf{R}^v\}_{v=1}^M$ from different views in the same sample are similar. Hence, the global representation $\mathbf{H}$ and view-consensus representations $\{\mathbf{R}^v\}_{v=1}^M$ from different views of the same sample should be mapped close together. In this respect, we denote $\{\mathbf{H}_i, \mathbf{R}_j^v\}_{j=i}^{v=1,\ldots,M}$ as M positive feature pairs, and the rest $\{\mathbf{H}_i, \mathbf{R}_j^v\}_{j\neq i}^{v=1,\ldots,M}$ are $M(N-1)$ negative feature pairs. To implement multi-view contrastive fusion, we first use the cosine distance to measure the similarity of feature pairs:

$$sim\left(\mathbf{H}_i, \mathbf{R}_j^v\right) = \frac{\left\langle \mathbf{H}_i, \mathbf{R}_j^v \right\rangle}{\|\mathbf{H}_i\| \left\|\mathbf{R}_j^v\right\|}, \tag{6}$$

where $\langle\cdot,\cdot\rangle$ is the dot product operator. We introduce a temperature parameter τ to moderate the effect of similarity, and $\mathbb{1}[j \neq i] \in \{0,1\}$ denotes the indicator function. For the v-th view, the contrastive fusion maximizes the similarities of positive pairs, and minimizes that of negative pairs:

$$\mathcal{L}_{\boldsymbol{CL}}^v\left(\mathbf{R}^v, \mathbf{H}\right) = -\frac{1}{N}\sum_{i=1}^N \log \frac{e^{sim\left(\mathbf{H}_i, \mathbf{R}_i^v\right)/\tau}}{\sum_{j=1}^N \mathbb{1}\left[j \neq i\right] e^{sim\left(\mathbf{H}_i, \mathbf{R}_j^v\right)/\tau}}. \tag{7}$$

Self-weighting method: The characteristics and qualities inherent in different views typically exhibit significant variation. In most previous works, such as [11], [24], multi-view contrastive learning is applied in an equal-sum manner, *e.g.*, $\sum_{m,n} \mathcal{L}_{CL}^{m,n}(\mathbf{R}^m, \mathbf{R}^n)$. Intuitively, high-quality views would be forced to align with low-quality views during contrastive learning to achieve maximum consistency. To mitigate it, we encourage conducting the consistency objective in a self-weighting manner, *i.e.*, $\sum_v \mathcal{W}^v \mathcal{L}_{\boldsymbol{CL}}^v(\mathbf{R}^v, \mathbf{H})$. Here, $\mathcal{W}^v$ adaptively adjusts the weight of each view in feature fusion. Concretely, if the view is useful and with informative semantics, contrastive learning between them is adaptively strengthened. Conversely, for unreliable views, contrastive learning between them is adaptively weakened. In this manner, high-quality views will dominate the feature fusion process, significantly mitigating the representation degeneration problem. In light of this, we redefine multi-view contrastive loss as:

$$\mathcal{L}_{\boldsymbol{CL}} = \sum_{v=1}^M \mathcal{W}^v \mathcal{L}_{\boldsymbol{CL}}^v\left(\mathbf{R}^v, \mathbf{H}\right), \tag{8}$$

where $\mathcal{W}^v$ is an adaptive weight between the global representation $\mathbf{H}$ and view-consensus representation $\mathbf{R}^v$.

In the unsupervised case, it is difficult to distinguish which representations in $\{\mathbf{R}^v\}_{v=1}^M$ are meaningless noise, and which contain valuable semantic information. To simplify it, we propose to measure the discrepancy between global features $\mathbf{H}$ and view-consensus features $\mathbf{R}$. The features $\mathbf{R}^v$ having a lower discrepancy with global features $\mathbf{H}$ have higher correlation, and are consequently assigned higher view weights, *i.e.*, $\mathcal{W}^v$. To this end, we define $\mathcal{D}(\mathbf{R}^v, \mathbf{H})$ as the discrepancy between $\mathbf{R}^v$ and $\mathbf{H}$, and denote $\mathcal{P}(\cdot)$ as a weight decision function. The view weight is updated by:

$$\mathcal{W}^v = \mathcal{P}\left(\mathcal{D}\left(\mathbf{R}^v, \mathbf{H}\right)\right). \tag{9}$$

To estimate the correlation among different feature pairs, the maximum mean discrepancy (**MMD**) [39] can effectively measure the discrepancy between two distributions $\mathbf{P}$ and $\mathbf{Q}$ based on the expectations of the two view data $\mathbf{X}_s = \{\mathbf{X}_i^s\}_{i=1}^{n_s}$ and $\mathbf{Y}_t = \{\mathbf{Y}_j^t\}_{j=1}^{n_t}$. The $\mathbf{X}_s$ and $\mathbf{Y}_t$ are generated from distributions $\mathbf{P}$ and $\mathbf{Q}$, respectively. Mathematically, **MMD** can be expressed as:

$$\mathbf{MMD}\left(\mathbf{X}_s, \mathbf{Y}_t\right) = \left\| \frac{1}{n_s}\sum_{i=1}^{n_s} \phi\left(\mathbf{X}_i^s\right) - \frac{1}{n_t}\sum_{j=1}^{n_t} \phi\left(\mathbf{Y}_j^t\right) \right\|_{\mathbb{H}}, \tag{10}$$

where $\mathbb{H}$ represents a Reproducing Kernel Hilbert Space (RHKS), and $\phi(\cdot)$ is the nonlinear feature mapping function (*e.g.*, Gaussian kernel). Then, getting square on both sides:

$$\begin{aligned}\mathbf{MMD}^2 &= \left\| \frac{1}{n_s}\sum_{i=1}^{n_s} \phi\left(\mathbf{X}_i^s\right) - \frac{1}{n_t}\sum_{j=1}^{n_t} \phi\left(\mathbf{Y}_j^t\right) \right\|_{\mathbb{H}}^2 \\ &= \left\| \frac{1}{n_s}\sum_{i=1}^{n_s} \phi\left(\mathbf{X}_i^s\right) \right\|_{\mathbb{H}}^2 + \left\| \frac{1}{n_t}\sum_{j=1}^{n_t} \phi\left(\mathbf{Y}_j^t\right) \right\|_{\mathbb{H}}^2 \\ &\quad - 2\left\| \frac{1}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} \phi\left(\mathbf{X}_i^s\right)\phi\left(\mathbf{Y}_j^t\right) \right\|_{\mathbb{H}}. \end{aligned} \tag{11}$$

In a Reproducing Kernel Hilbert Space, $k(\mathbf{X}_i^s, \mathbf{Y}_j^t)$ denotes the inner product of $\phi(\mathbf{X}_i^s)$ and $\phi(\mathbf{Y}_j^t)$. We expand (11), and $\mathbf{MMD}^2$ finally can be formulated as:

$$\begin{aligned}\mathbf{MMD}^2\left(\mathbf{X}_s, \mathbf{Y}_t\right) &= \frac{1}{{n_s}^2}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s} k\left(\mathbf{X}_i^s, \mathbf{X}_j^s\right) \\ &\quad + \frac{1}{{n_t}^2}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t} k\left(\mathbf{Y}_i^t, \mathbf{Y}_j^t\right) \\ &\quad - \frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} k\left(\mathbf{X}_i^s, \mathbf{Y}_j^t\right). \end{aligned} \tag{12}$$

In the end, the **MMD** is implemented to estimate the discrepancy between the features $\mathbf{R}^v$ and global features $\mathbf{H}$, *i.e.*, $\mathcal{D}(\mathbf{R}^v, \mathbf{H})$. In particular, MMD is a non-parametric method that avoids specific assumptions about the distribution's form, allowing it applicable to a diverse range of data types. We use linear kernel (*i.e.*, $k(x,y) = x^T y$) to project representations into

RKHS. Since the features $\mathbf{R}^v$ have the same size as global features $\mathbf{H}$, the $\mathcal{D}(\mathbf{R}^v, \mathbf{H})$ can be expressed as:

$$\begin{aligned}\mathcal{D}_{MMD}\left(\mathbf{R}^v, \mathbf{H}\right) = &\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k\left(\mathbf{R}_i^v, \mathbf{R}_j^v\right) \\ &+ \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k\left(\mathbf{H}_i, \mathbf{H}_j\right) \\ &- \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k\left(\mathbf{R}_i^v, \mathbf{H}_j\right). \end{aligned} \tag{13}$$

Considering that the features $\mathbf{R}^v$ having lower discrepancy with global features $\mathbf{H}$ should be weighted a higher value, we employ a normalized weight decision function $\mathcal{P}(\mathcal{D}(\mathbf{R}^v, \mathbf{H})) = Softmax(-\mathcal{D}(\mathbf{R}^v, \mathbf{H}))$ for computing the weight $\mathcal{W}^v$. The final view weight is updated by:

$$\mathcal{W}^v = \mathcal{P}\left(\mathcal{D}\left(\mathbf{R}^v, \mathbf{H}\right)\right) = \frac{e^{-\mathcal{D}_{MMD}\left(\mathbf{R}^v, \mathbf{H}\right)}}{\sum_{v'} e^{-\mathcal{D}_{MMD}\left(\mathbf{R}^{v'}, \mathbf{H}\right)}}. \tag{14}$$

In consequence, the view-consensus features can be written as $\mathbf{R}^v = \mathcal{R}(E(\mathbf{X}^v))$, allowing them to filter out the view-private information of $\{\mathbf{Z}^v\}_{v=1}^M$. And the global features can be written as $\mathbf{H} = \mathcal{F}([\mathbf{Z}^1, \mathbf{Z}^2, \ldots, \mathbf{Z}^M])$, fully exploring cross-view consensus and complementary information. More importantly, $\mathcal{W}^v$ adaptively adjusts the consistency objective between $\mathbf{H}$ and $\{\mathbf{R}^v\}_{v=1}^M$, where the useful views will dominant feature fusion while the unreliable views are weakened, significantly mitigating representation degeneration. Overall, the loss of our proposed SCMVC is:

$$\begin{aligned}\mathcal{L}_{\boldsymbol{total}} &= \mathcal{L}_{\boldsymbol{Z}} + \mathcal{L}_{\boldsymbol{CL}} \\ &= \mathcal{L}_{\boldsymbol{Z}}\left(\left\{\mathbf{X}^v, \hat{\mathbf{X}}^v\right\}_{v=1}^M; \{\theta^v, \phi^v\}_{v=1}^M\right) \\ &\quad + \mathcal{L}_{\boldsymbol{CL}}\left(\{\mathbf{R}^v, \mathbf{H}\}_{v=1}^M; \Psi, \Phi, \{\theta^v\}_{v=1}^M\right), \end{aligned} \tag{15}$$

where $\mathcal{L}_{\boldsymbol{Z}}$ and $\mathcal{L}_{\boldsymbol{CL}}$ are the reconstruction and consistency objectives that are conducted in different feature spaces. Thanks to our self-weighting method, we do not need weight parameters to balance the different losses, *i.e*, decreasing the hyperparameter λ in (5).

C. Clustering Module

For the final clustering task, we take the k-means algorithm on global features $\mathbf{H}$ to obtain the clustering results for all samples. Specifically, the learned global representation $\mathbf{H}$ is factorized as follows:

$$\begin{aligned}&\min_{\mathbf{U},\mathbf{C}} \|\mathbf{H} - \mathbf{U}\mathbf{C}\|^2 \\ &\text{s.t. } \mathbf{U}\mathbf{1} = \mathbf{1}, \mathbf{U} \geq \mathbf{0}\end{aligned} \tag{16}$$

where $\mathbf{U} \in \mathbb{R}^{N\times K}$ is cluster indicator matrix. $\mathbf{C} \in \mathbb{R}^{K\times d_h}$ is the center matrix of clustering.

TABLE II
SUMMARY OF DATASETS USED IN OUR STUDY

Datasets	Samples	Views	Classes
MNIST-USPS [40]	5,000	2	10
BDGP [41]	2,500	2	5
Prokaryotic [42]	551	3	4
Synthetic3d [43]	600	3	3
CCV [44]	6,773	3	20
Fashion [45]	10,000	3	10
Cifar10[1]	50,000	3	10
Cifar100[1]	50,000	3	100
Caltech-2V [46]	1,400	2	7
Caltech-3V [46]	1,400	3	7
Caltech-4V [46]	1,400	4	7
Caltech-5V [46]	1,400	5	7

Algorithm 1: Optimization of SCMVC.

Input: Multi-view dataset $\{\mathbf{X}^v\}_{v=1}^M$; Cluster number k.
Output: Cluster indicator matrix $\mathbf{U}$.

```
Pretrain autoencoders {θ^v, φ^v}_{v=1}^M by minimizing Eq. (1).
while Not reach the maximum iteration T_max
    Update the view weight by Eqs. (13,14).
    Calculate the multi-view contrastive loss by Eq. (8).
    Compute the overall loss function by Eq. (15).
    Back propagation and train the whole SCMVC model.
end while
Compute U on the final global features H by Eq. (16).
```

D. Optimization

The entire optimization process of SCMVC is summarized in Algorithm 1. The model consists of multiple autoencoder models and two MLPs. Specifically, we adopt the mini-batch gradient descent algorithm to optimize the model. First, all autoencoders are initialized by (1). Second, the self-weighted contrastive fusion is conducted to achieve the consistency objective by (3). In the end, we compute the global representation by (4), and the cluster indicator matrix can be obtained by (16).

IV. EXPERIMENTS

A. Experimental Setup

1) Multi-View Datasets Description: To comprehensively evaluate the performance of our proposed model SCMVC, we conduct experiments on nine publicly available multi-view datasets, which are shown in Table II. Specifically, there are four small-scale multi-view datasets, including MNIST-USPS [40], BDGP [41], Prokaryotic [42], Synthetic3d [43] to validate the effectiveness of SCMVC in multi-view tasks. Furthermore, to further explore the model generalization, we select four large-scale datasets, *i.e.*, CCV [44], Fashion [45], Cifar10,[1] Cifar100[1]. In the end, we build four datasets based on Caltech [46] that "Caltech-*X*V" represents that it consists of X views, allowing

[1]http://www.cs.toronto.edu/kriz/cifar.html

TABLE III
RESULTS FOR ALL METHODS ON FOUR SMALL DATASETS

Small-scale datasets	MNIST-USPS			BDGP			Prokaryotic			Synthetic3d		
Evaluation metrics	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
LMVSC [48] (2020)	0.672	0.507	0.675	0.821	0.842	0.917	0.565	0.129	0.619	0.957	0.831	0.957
CGD [47] (2020)	0.990	0.975	0.990	0.730	0.657	0.754	0.565	0.298	0.646	0.790	0.443	0.790
DEMVC [49] (2021)	0.996	0.989	0.996	0.751	0.750	0.751	0.535	0.316	0.693	0.832	0.654	0.832
CoMVC [11] (2021)	0.987	0.977	0.987	0.812	0.733	0.813	0.424	0.213	0.675	0.953	0.818	0.952
CONAN [12] (2021)	0.553	0.563	0.553	0.968	0.930	0.968	0.491	0.163	0.505	0.965	0.854	0.965
EOMSC [16] (2022)	0.765	0.660	0.765	0.943	0.873	0.943	0.548	0.312	0.655	0.963	0.851	0.963
MFLVC [9] (2022)	0.993	0.985	0.993	0.986	0.960	0.986	0.472	0.256	0.587	0.965	0.854	0.965
DSMVC [24] (2022)	0.723	0.729	0.733	0.658	0.444	0.658	0.456	0.170	0.592	0.948	0.802	0.948
GCFAggMVC [27] (2023)	0.995	0.986	0.995	0.987	0.961	0.987	0.623	0.378	0.731	0.970	0.871	0.970
DealMVC [50] (2023)	0.922	0.910	0.922	0.988	0.963	0.988	0.632	0.278	0.632	0.877	0.657	0.877
SCMVC (ours)	**0.997**	**0.990**	**0.997**	**0.992**	**0.975**	**0.992**	**0.742**	**0.472**	**0.842**	**0.982**	**0.918**	**0.982**

Bold denotes the best results and underline denotes the second best results.

us to investigate our model robustness as the number of views increases.

2) Comparing Methods: To demonstrate the performance of our proposed SCMVC, we compare it with 10 state-of-the-art multi-view clustering methods:

Shallow multi-view clustering methods: **CGD**: multi-view clustering via cross-view graph diffusion [47], **LMVSC**: large-scale multi-view subspace clustering [48], and **EOMSC**: efficient one-pass multi-view subspace clustering [16].

Deep multi-view clustering methods: **DEMVC**: deep embedded multi-view clustering with collaborative training [49], **CoMVC**: contrastive multi-view clustering [11], **CONAN**: contrastive fusion networks for multi-view clustering [12], **MFLVC**: multi-level feature learning for contrastive multi-view clustering [9], **DSMVC**: deep safe multi-view clustering [24], **GCFAggMVC**: global and cross-view feature aggregation for multi-view clustering [27], and **DealMVC**: dual contrastive calibration for multi-view clustering [50]. Among them, **CoMVC**, **CONAN**, **MFLVC**, **GCFAggMVC**, and **DealMVC** employ multi-view contrastive learning to implement the consistency objective.

3) Evaluation Metrics: Three widely used metrics are applied to evaluate clustering performance, *i.e.*, clustering accuracy (ACC), normalized mutual information (NMI), and purity (PUR). The higher the values of these metrics, the better the clustering results. The mean values of 10 runs are reported for all multi-view clustering methods.

4) Implementation Details: All datasets are reshaped into vectors, and the fully connected (Fc) autoencoders with a similar architecture are used for extracting low-level features $\{\mathbf{Z}^v\}_{v=1}^M$. Specifically, for each view, the structure of the encoder is: Input - Fc500 - Fc500 - Fc2000 - Fc64, and the decoder is symmetric with the encoder. After that, we use a linear MLP, which is constructed as Input(64) - Fc20, to extract view-consensus features $\{\mathbf{R}^v\}_{v=1}^M$, and another non-linear MLP with two-layer architecture, *i.e.*, Input($M\times$64) - Fc256 - Fc20, to learn global features $\mathbf{H}$. The following settings are the same for all experimental datasets. The ReLU activation function is used in all the layers except for the output layer. Adam is chosen as the optimizer with a default learning rate of 0.0003. The experiments are conducted on a Windows PC with Intel (R) Core (TM) i5-9300H CPU@2.40 GHz, 16.0 GB RAM, and TITAN X GPU (12 GB caches).

B. Comparative Result Analysis

The comparison results on eight datasets are shown in Tables III and IV. We can observe that the proposed SCMVC achieves the best results than the previous MVC methods. Particularly, we have the following observations: (1) Comparing three shallow multi-view clustering methods (*i.e.*, GCD, LMVSC, and EOMSC), we can find that these methods attempt to learn the data subspace representation or graph structure relationship from raw data, whereas much meaningless private information is retained. This information is harmful to learn latent features, even causing model collapse. (2) Comparing two traditional deep multi-view clustering methods (*i.e.*, DEMVC and DSMVC), we can find that these methods implement autoencoder models to learn the salient representation from raw data, where the consistency and reconstruction objectives are retained in the same feature space. The meaningless view-private information is constantly reconstructed, which in turn produces suboptimal solutions.

Moreover, our method outperforms previous contrastive multi-view clustering methods (*i.e.*, CONAN, CoMVC, MFLVC, GCFAggMVC, and DealMVC). Specifically, we find that: (3) In CoMVC and MFLVC, they explore view consistency by aligning the representations of each view, which has a negative impact since the lack of global complementary information is prone to produce inferior solutions. Taking the Prokaryotic dataset for example, our proposed SCMVC improves ACC by 27.0% and 31.8% compared to MFLVC and CoMVC, respectively. (4) In CONAN and GCFAggMVC, they obtain a consensus feature representation by contrastive fusion way. However, they treat each view equally, where the low-quality views may dominate the entire feature fusion process, and thus harmful to clustering. Taking the CCV dataset for example, our proposed SCMVC improves ACC by 5.2% compared to the second-best baseline GCFAggMVC. In conclusion, our proposed SCMVC method enhances view complementarity through global information aggregation, and emphasizes the importance

TABLE IV
RESULTS FOR ALL METHODS ON FOUR LARGE DATASETS

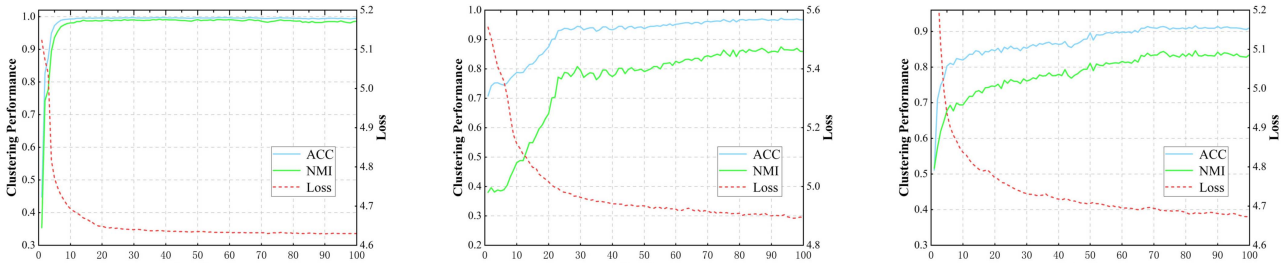
Large-scale datasets	CCV			Fashion			Cifar10			Cifar100		
Evaluation metrics	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
LMVSC [48] (2020)	0.207	0.167	0.249	0.769	0.732	0.790	0.988	0.972	0.988	0.848	0.966	0.958
CGD [47] (2020)	0.167	0.124	0.204	0.802	0.817	0.802	0.967	0.961	0.971	0.825	0.951	0.829
DEMVC [49] (2021)	0.162	0.119	0.173	0.827	0.856	0.827	0.435	0.366	0.451	0.524	0.705	0.559
CoMVC [11] (2021)	0.295	0.287	0.311	0.867	0.872	0.881	0.926	0.892	0.926	0.657	0.935	0.657
CONAN [12] (2021)	0.152	0.121	0.168	0.827	0.831	0.776	0.928	0.894	0.928	0.673	0.927	0.714
EOMSC [16] (2022)	0.221	0.177	0.236	0.757	0.784	0.759	0.836	0.735	0.836	0.583	0.923	0.584
MFLVC [9] (2022)	0.298	0.307	0.332	0.992	0.980	0.992	0.991	0.977	0.991	0.826	0.943	0.826
DSMVC [24] (2022)	0.179	0.144	0.210	0.778	0.780	0.783	0.721	0.649	0.722	0.730	0.953	0.730
GCFAggMVC [27] (2023)	0.348	0.324	0.373	0.895	0.938	0.903	0.992	0.978	0.992	0.959	0.993	0.961
DealMVC [50] (2023)	0.311	0.302	0.320	0.895	0.955	0.895	0.969	0.971	0.969	0.852	0.967	0.892
SCMVC (ours)	**0.400**	**0.336**	**0.415**	**0.995**	**0.986**	**0.995**	**0.995**	**0.985**	**0.995**	**0.983**	**0.995**	**0.987**

Bold denotes the best results and underline denotes the second best results.

TABLE V
RESULTS FOR ALL MULTI-VIEW CLUSTERING METHODS ON CALTECH DATASET

Datasets	Caltech-2V			Caltech-3V			Caltech-4V			Caltech-5V		
Evaluation metrics	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
LMVSC [48] (2020)	0.560	0.432	0.579	0.704	0.589	0.708	0.759	0.678	0.768	0.765	0.672	0.770
CGD [47] (2020)	0.474	0.438	0.544	0.690	0.599	0.724	0.744	0.674	0.777	0.771	0.726	0.803
DEMVC [49] (2021)	0.486	0.342	0.486	0.505	0.366	0.512	0.454	0.315	0.474	0.457	0.378	0.489
EOMSC [16] (2022)	0.466	0.343	0.474	0.531	0.390	0.534	0.565	0.433	0.579	0.596	0.490	0.596
DSMVC [24] (2022)	0.603	0.526	0.619	0.745	0.674	0.745	0.834	**0.766**	0.834	**0.919**	0.847	**0.919**
CoMVC [11] (2021)	0.462	0.417	0.503	0.543	0.511	0.584	0.583	0.527	0.614	0.667	0.576	0.697
CONAN [12] (2021)	0.565	0.442	0.565	0.597	0.502	0.597	0.622	0.541	0.657	0.721	0.642	0.721
MFLVC [9] (2022)	0.606	0.528	0.616	0.633	0.561	0.633	0.717	0.624	0.728	0.797	0.691	0.798
GCFAggMVC [27] (2023)	0.644	0.512	0.644	0.640	0.535	0.640	0.734	0.661	0.734	0.821	0.703	0.821
DealMVC [50] (2023)	0.606	0.508	0.606	0.644	0.575	0.673	0.823	0.726	0.823	0.870	0.780	0.870
SCMVC (ours)	**0.674**	**0.533**	**0.674**	**0.759**	**0.663**	**0.759**	**0.844**	0.729	**0.844**	**0.919**	**0.867**	**0.919**

"-XV" represents that it consists of X views.
Bold denotes the best results and underline denotes the second best results.



(a) Loss *vs* Performance (MNIST-USPS) (b) Loss *vs* Performance (Syntheic3d) (c) Loss *vs* Performance (Caltech-5V)

Fig. 4. (a–c) Training process analysis (*i.e.*, loss and performance variations) on MNIST-USPS, Syntheic3d, and Caltech-5 V, respectively.

of high-quality views in feature fusion to learn the global features more effectively.

To further validate the robustness of our SCMVC method as the number of views increases, we test the performance of different numbers of views on the Caltech dataset. Table V shows the comparative results with the selected competitors. We can observe that our proposed SCMVC is more robust compared to the previous MVC methods. This is because our proposed self-weighting method in (8) can adaptively strengthen useful views, and can reduce unreliable views, which significantly mitigates global features to discard useful semantics, and thus exhibit strong robustness.

C. Model Analysis

1) Visualization of the Clustering Results: In order to visually investigate the effectiveness of the proposed SCMVC, the t-SNE algorithm [51] is applied to visualize the distribution of latent embedding of different levels, *i.e.*, the features $\mathbf{Z}$, $\mathbf{R}$, and $\mathbf{H}$. As shown in Fig. 5, the clusters of global features $\mathbf{H}$ are clearer than low-level features $\mathbf{Z}$ and view-consensus features $\mathbf{R}$, exhibiting a denser cluster structure. These results all confirm the effectiveness of SCMVC.

2) Convergence Analysis: It is not difficult to discover that the reconstruction objective $\mathcal{L}_Z$ and the consistency objective

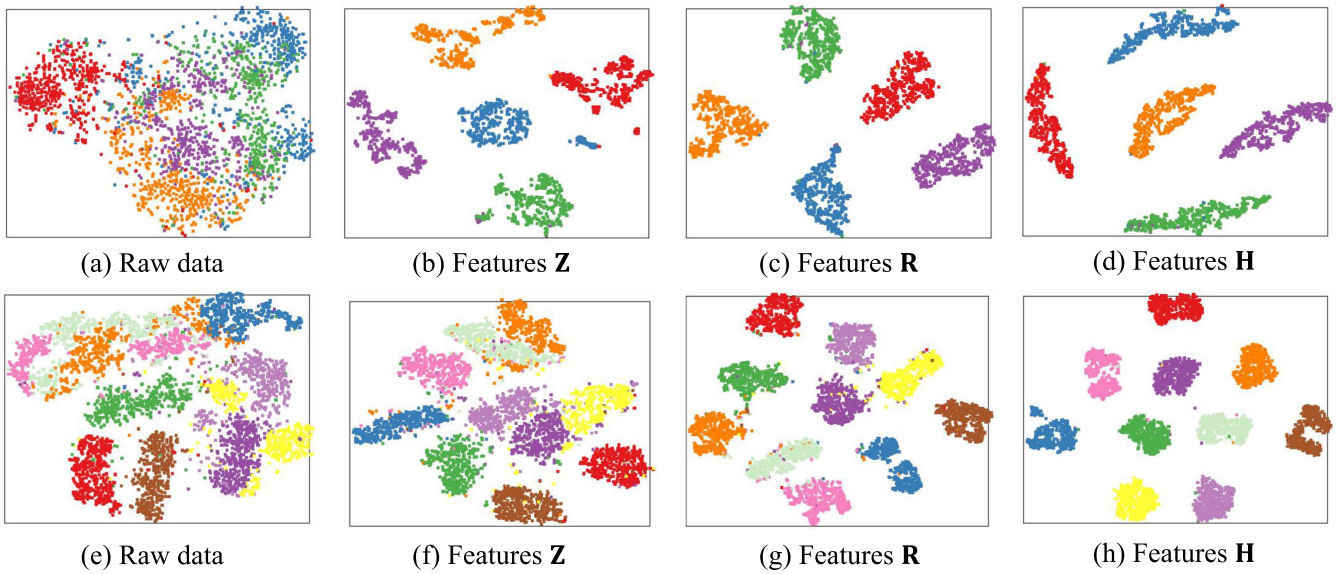


Fig. 5. (a–d) Visualization results on BDGP dataset. (e-h) Visualization results on MNIST-USPS dataset. Specifically, the features $\mathbf{Z}$, $\mathbf{R}$, and $\mathbf{H}$ denote low-level features, view-consensus features, and global features, respectively.

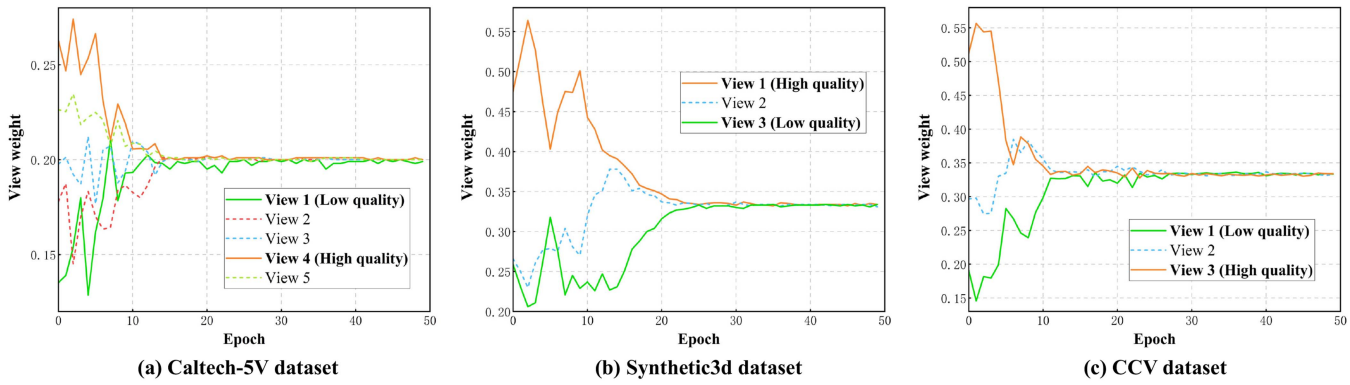


Fig. 6. View weighting analysis on Caltech-5 V, Synthetic3d and CCV datasets, respectively. Different views are firstly given different weights, and as the number of iterations increases, the weights of the different views tend to converge.

$\mathcal{L}_{CL}$, *i.e.*, ((1), (7)) are all convex functions. As shown in Fig. 4, it can be observed that the loss value monotonically decreases until reaching convergence, while the values of ACC and NMI exhibit an initial gradual increase followed by fluctuating within a narrow range. These results confirm the convergence of SCMVC.

3) Parameter Sensitivity Analysis: Thanks to our well-designed self-weighted contrastive fusion framework, we do not need numerous hyperparameters to balance different loss compositions. Specifically, in this section, we explore the best settings of hyperparameters τ for (7). Fig. 8 demonstrates the ACC, NMI, and PUR of SCMVC when the hyperparameter τ is tuned in the range of $\{0.1, 0.3, 0.5, 0.7, 1\}$. We could observe that: (1) When τ is at a small value, the clustering performance of the proposed SCMVC decreases. This may be because the excessive pursuit of view consistency might result in intrinsic feature space being inseparable. (2) When the value of τ increases, the clustering performance gradually recovers, and they are insensitive to τ in the range 0.5 to 1. Empirically, we set $\tau = 0.3$ for the CCV dataset, $\tau = 0.5$ for all Caltech datasets, and $\tau = 1$ for other multi-view datasets.

4) View Weighting Analysis: The self-weighting method is one of key components in our SCMVC, which adaptively strengthens useful views in feature fusion, and weakens unreliable views. In this section, we further explore how the self-weighting method adjusts the multi-view contrastive learning. Specifically, Fig. 6 illustrates the change of weights with iterations for different views on Caltech-5 V, Prokaryotic, and CCV datasets, respectively. We can find that: (1) Initially, varying weights are assigned to different views. The high-quality views are weighted by higher values and low-quality views are de-weighted. Correspondingly, contrastive learning with high-quality views will be strengthened, while mitigating the loss caused by aligning with low-quality views, which is indicated in Fig. 7. (2) With the increase in the number of iterations, the weights of different views gradually converge. This is because multi-view contrastive learning is capable of rapidly closing the

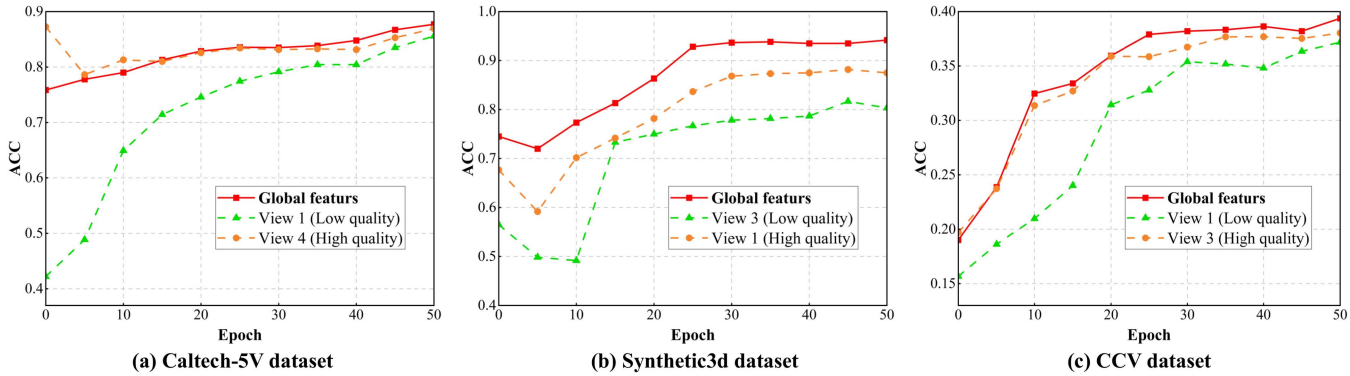


Fig. 7. Variation in ACC for different quality views with the iterations of contrastive learning on Caltech-5 V, Synthetic3d and CCV datasets, respectively. The global features can learn reliable semantics from high-quality views, while reducing the impact of low-quality views.

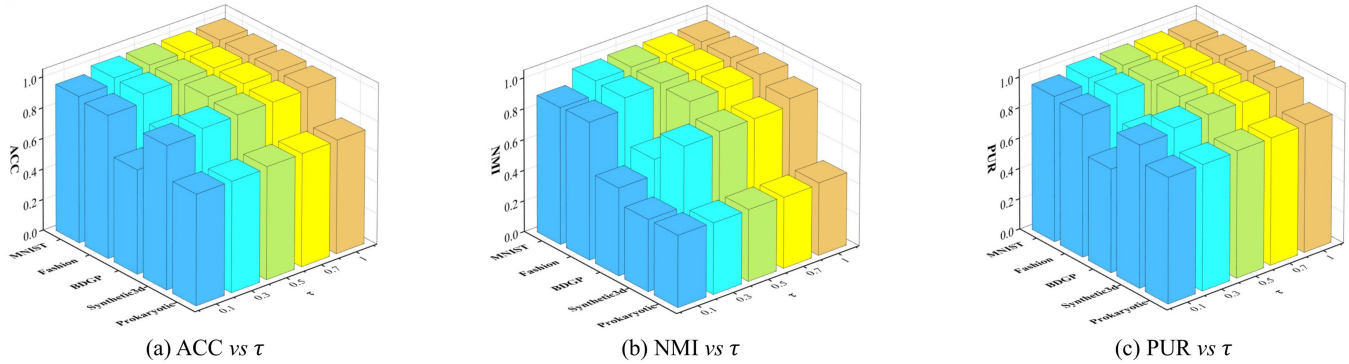


Fig. 8. (a-c) Parameter sensitivity analysis on five multi-view datasets, including MNIST-USPS, Fashion, BDGP, Synthetic3d, and Prokaryotic.

semantic gap among different views, where the discrepancy between view-consensus features $\mathbf{R}^v$ and global features $\mathbf{H}$ becomes progressively consistent.

Taking the Caltech-5 V dataset as an example, as depicted in Fig. 1(b), we can observe that view 4 is the high-quality view, while view 1 is the low-quality view. Correspondingly, our self-weighting method gives a higher weight for view 4, while de-weighting view 1 in Fig. 6(a). In this way, global features can be better aligned with high-quality view 4, while contrastive learning with low-quality view 1 is adaptively weakened. As shown in Fig. 7(a), global features can effectively keep consistent with high-quality views, thereby learning more useful semantics from reliable views. Finally, the weights of view 1 and view 4 gradually converge to the mean value with iterations increases to achieve the consistency objective. Notably, as shown in Fig. 7(b), global features benefiting from complementary information tend to achieve superior clustering performance. These results demonstrate the effectiveness of our proposed self-weighting multi-view contrastive fusion method.

D. Ablation Studies

1) Loss Components: To understand the effectiveness of the proposed SCMVC components, we remove each component individually to observe the change in performance. Specifically, (A) **MCL** denotes the multi-view contrastive learning to implement the consistency objective. (B) **SEW** denotes a self-weighting method to adaptively weight each view. (C) **SCMVC** denotes the complete multi-view contrastive fusion of our method. As shown in Table VI, **MCL**, *i.e.*, $\mathcal{L}_{CL}$, plays a crucial role in SCMVC, and without it, the model performance shows a decrease of 10.9%, 20.0%, and 14.3% in terms of ACC on the CCV, Caltech-5 V, and Prokaryotic datasets, respectively. This is because global features $\mathbf{H}$ computed without $\mathcal{L}_{CL}$ are disturbed by the inherent irrelevant information from each view, which severely impacts clustering performance. Furthermore, **SEW** could further optimize the entire multi-view contrastive fusion framework, where the model performance improves by

TABLE VI
ABLATION STUDIES ON LOSS COMPONENTS

Datasets	Method	ACC	NMI	PUR
CCV	No-MCL	0.194	0.156	0.236
	No-SEW	0.303	0.296	0.303
	SCMVC	**0.400**	**0.336**	**0.415**
Caltech-5V	No-MCL	0.663	0.585	0.689
	No-SEW	0.863	0.782	0.863
	SCMVC	**0.919**	**0.867**	**0.919**
Prokaryotic	No-MCL	0.490	0.256	0.686
	No-SEW	0.633	0.365	0.722
	SCMVC	**0.742**	**0.472**	**0.842**

Bold denotes the best results.

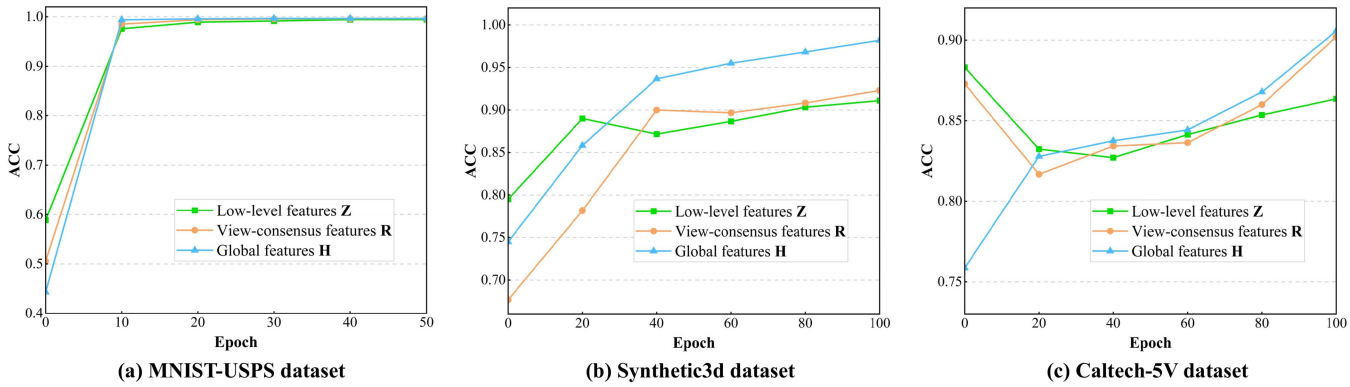


Fig. 9. Clustering performance for low-level features $\mathbf{Z}$, view-consensus features $\mathbf{R}$, and global features $\mathbf{H}$ on three small-scale multi-view datasets.

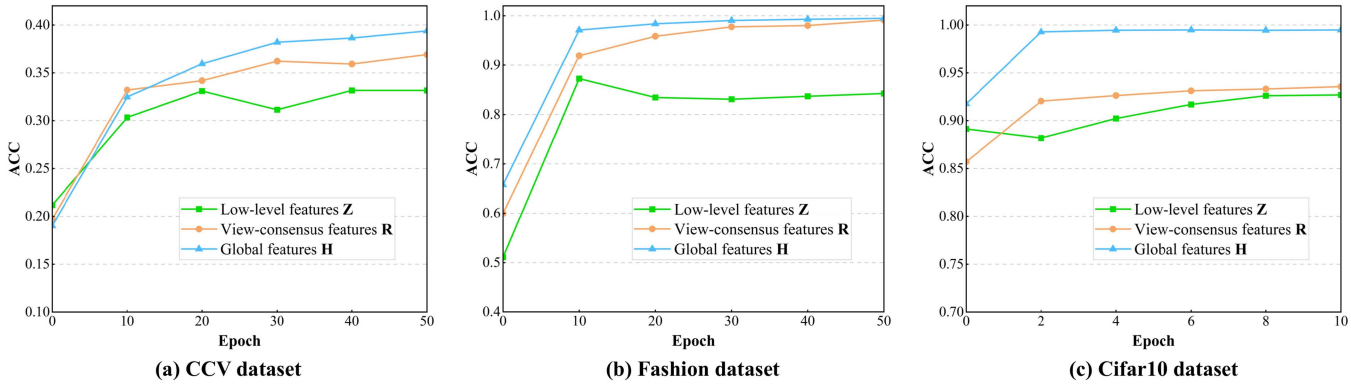


Fig. 10. Clustering performance for low-level features $\mathbf{Z}$, view-consensus features $\mathbf{R}$, and global features $\mathbf{H}$ on three large-scale multi-view datasets.

9.7%, 5.6%, and 10.9% in ACC on the CCV, Caltech-5 V, and Prokaryotic datasets, respectively. These results confirm the effectiveness of the proposed **MCL** and **SEW**.

2) Hierarchical Feature Fusion Framework: In our SCM-VC, we introduce a hierarchical feature fusion framework where different objectives are conducted in different feature spaces. Compared to the previous MCL framework like Fig. 1(a), our framework further explores cross-view complementary information via feature fusion. Meanwhile, the consistency objective is redesigned to implement between view-consensus features and global features. To further verify the superiority of our architecture, we perform k-means algorithms on different level features, as shown in Figs. 9 and 10. In particular, for the results with multiple views, we select their best one. The results indicate that our hierarchical feature fusion framework can make high-level features, *i.e.*, view-consensus features and global features, to capture more reliable semantic information. Meanwhile, global features exhibit optimal performance for downstream clustering tasks.

V. Conclusion

In this paper, we propose a novel framework of self-weighted contrastive fusion for deep multi-view clustering, where the consistency objective is effectively separated from the reconstruction objective. To fully explore view consistency and complementarity, we maximize consensus expression between global features, which summarizes the global common information of each view, and view-consensus features. Particularly, a self-weighting method is introduced to adaptively strengthen useful views in feature fusion, and weaken unreliable views, significantly mitigating the representation degeneration problem. Extensive experimental results on nine public datasets demonstrate that our proposed method outperforms state-of-the-art multi-view clustering methods.

References

[1] 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.
[2] Y. Wang, D. Chang, Z. Fu, and Y. Zhao, "Consistent multiple graph embedding for multi-view clustering," *IEEE Trans. Multimedia*, vol. 25, pp. 1008–1018, 2023.
[3] Y. Chen, X. Zhang, J. Wang, and J. Li, "Large-scale multi-view clustering based on anchor strategy and tensor collaborative learning," in *Proc. IEEE 9th Int. Conf. Cloud Comput. Intell. Syst.*, 2023, pp. 18–23.
[4] Y. Chen et al., "Self-paced enhanced low-rank tensor kernelized multi-view subspace clustering," *IEEE Trans. Multimedia*, vol. 24, pp. 4054–4066, 2022.
[5] C. Xu et al., "Deep multi-view concept learning," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 2898–2904.
[6] B. Cui, H. Yu, L. Zong, and Z. Cheng, "Self-guided deep multi-view subspace clustering network," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[7] F. Nie, G. Cai, and X. Li, "Multi-view clustering and semi-supervised classification with adaptive neighbours," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 2408–2414.
[8] S. Wei, J. Wang, G. Yu, C. Domeniconi, and X. Zhang, "Deep incomplete multi-view multiple clusterings," in *Proc. IEEE Int. Conf. Data Mining*, 2020, pp. 651–660.
[9] J. Xu et al., "Multi-level feature learning for contrastive multi-view clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16051–16060.
[10] Z. Huang, Y. Ren, X. Pu, and L. He, "Non-linear fusion for self-paced multi-view clustering," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 3211–3219.
[11] D. J. Trosten, S. Lokse, R. Jenssen, and M. Kampffmeyer, "Reconsidering representation alignment for multi-view clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1255–1265.
[12] G. Ke et al., "CONAN: Contrastive fusion networks for multi-view clustering," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, 2021, pp. 653–660.
[13] 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.
[14] C. Fettal, L. Labiod, and M. Nadif, "Scalable attributed-graph subspace clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 7559–7567.
[15] C. Zhang et al., "Generalized latent multi-view subspace clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 86–99, Jan. 2020.
[16] S. Liu et al., "Efficient one-pass multi-view subspace clustering with consensus anchors," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 7576–7584.
[17] C. Zhang et al., "Multi-view clustering via deep matrix factorization and partition alignment," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 4156–4164.
[18] M. Zhang and K. Liu, "Rethinking symmetric matrix factorization: A more general and better clustering perspective," in *Proc. IEEE Int. Conf. Data Mining*, 2022, pp. 695–702.
[19] S. Wei, J. Wang, G. Yu, C. Domeniconi, and X. Zhang, "Multi-view multiple clusterings using deep matrix factorization," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 6348–6355.
[20] Z. Lin and Z. Kang, "Graph filter-based multi-view attributed graph clustering," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, 2021, pp. 2723–2729.
[21] J. Chen et al., "Shared-attribute multi-graph clustering with global self-attention," in *Proc. Int. Conf. Neural Inf. Process.*, 2022, pp. 51–63.
[22] W. Xia, Q. Wang, Q. Gao, X. Zhang, and X. Gao, "Self-supervised graph convolutional network for multi-view clustering," *IEEE Trans. Multimedia*, vol. 24, pp. 3182–3192, 2022.
[23] S. Fan et al., "One2multi graph autoencoder for multi-view graph clustering," in *Proc. Web Conf.*, 2020, pp. 3070–3076.
[24] H. Tang and Y. Liu, "Deep safe multi-view clustering: Reducing the risk of clustering performance degradation caused by view increase," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 202–211.
[25] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 478–487.
[26] X. Guo, L. Gao, X. Liu, and J. Yin, "Improved deep embedded clustering with local structure preservation," in *Proc. 16th Int. Joint Conf. Artif. Intell.*, 2017, pp. 1753–1759.
[27] W. Yan et al., "GCFAGG: Global and cross-view feature aggregation for multi-view clustering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19863–19872.
[28] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[29] L. E. C. La Rosa and D. A. B. Oliveira, "Learning from label proportions with prototypical contrastive clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2153–2161.
[30] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Van Gool, "SCAN: Learning to classify images without labels," in *Proc. Eur. Conf. Computer Vis.*, 2020, pp. 268–285.
[31] Y. Li et al., "Contrastive clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 8547–8555.
[32] H. Zhong et al., "Graph contrastive clustering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9224–9233.
[33] S. Wang, X. Lin, Z. Fang, S. Du, and G. Xiao, "Contrastive consensus graph learning for multi-view clustering," *IEEE/CAA J. Automatica Sinica*, vol. 9, no. 11, pp. 2027–2030, Nov. 2022.
[34] Z. Xue et al., "Robust diversified graph contrastive network for incomplete multi-view clustering," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 3936–3944.
[35] H. Li, L. Zhang, and K. Su, "Dual mutual information constraints for discriminative clustering," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 8571–8579.
[36] G. E. Hinton and R. Zemel, "Autoencoders, minimum description length and Helmholtz free energy," in *Proc. 6th Int. Conf. Neural Inf. Process. Syst.*, 1993, pp. 3–10.
[37] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. 2nd Int. Conf. Learn. Representations*, 2014.
[38] K. He et al., "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.
[39] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel method for the two-sample-problem," in *Proc. 19th Int. Conf. Neural Inf. Process. Syst.*, 2006, pp. 513–520.
[40] X. Peng, Z. Huang, J. Lv, H. Zhu, and J. T. Zhou, "COMIC: Multi-view clustering without parameter selection," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 5092–5101.
[41] 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.
[42] M. Brbić et al., "The landscape of microbial phenotypic traits and associated genes," *Nucleic Acids Res.*, vol. 44, pp. 10074–10090, 2016.
[43] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 1413–1421.
[44] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui, "Consumer video understanding: A benchmark database and an evaluation of human and machine performance," in *Proc. 1st ACM Int. Conf. Multimedia Retrieval*, 2011, pp. 1–8.
[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.*, 2021, pp. 9234–9243.
[46] 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," *Comput. Vis. Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.
[47] C. Tang et al., "CGD: Multi-view clustering via cross-view graph diffusion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 5924–5931.
[48] Z. Kang et al., "Large-scale multi-view subspace clustering in linear time," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 4412–4419.
[49] J. Xu et al., "Deep embedded multi-view clustering with collaborative training," *Inf. Sci.*, vol. 573, pp. 279–290, 2021.
[50] X. Yang et al., "DealMVC: Dual contrastive calibration for multi-view clustering," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023, pp. 337–346.
[51] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

Song Wu is currently working toward the M.Sc. degree in computer science and technology with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include multiview clustering and computer vision.

Yan Zheng is currently working toward the M.Sc. degree in computer science and technology with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. Her research interests include unsupervised learning and e-Health.

Yazhou Ren (Member, IEEE) received the B.Sc. degree in information and computation science and the Ph.D. degree in computer science from the South China University of Technology, Guangzhou, China, in 2009 and 2014, respectively. He is currently an Associate Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. From 2012 to 2014, he visited the Data Mining Laboratory, George Mason University, Fairfax, VA, USA. He has authored or coauthored more than 100 papers in refereed conferences and journals, such as AAAI, IJCAI, NeurIPS, CVPR, ICCV, ECCV, ACM MM, SIGIR, ICDM, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, and *Information Fusion*. His research interests include multiview clustering, unsupervised learning, and medical data analysis.

Jing He received the Ph.D. degree from the Academy of Mathematics and System Science, Chinese Academy of Sciences, Beijing, China, in 2006. She is currently a Professor with University of Oxford, Oxford, U.K. She has authored or coauthored more than 150 research papers in refereed international journals and conference proceedings. Her research interests include data mining, web service/web search, spatial and temporal database, and multiple criteria decision making. Her research has attracted more than seven million research funds. In 2019, she has proposed a polynomial-time solution for graph isomorphism based on permutation and equinumerosity theorem (https://doi.org/10.1002/cpe.5484).

Xiaorong Pu received the Ph.D. degree in computer application from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2007. She is currently a Professor with the School of Computer Science and Engineering, UESTC, Chengdu, China. Her research interests include neural networks, computer vision, computer aided diagnosis, and e-Health.

Shudong Huang received the Ph.D. degree in computer science and engineering from the University of Electronic Science and Technology of China, Chengdu, China. He is currently an Associate Research Professor with the College of Computer Science, Sichuan University, Chengdu. His research interests include machine learning, representation learning, pattern recognition, and data mining.

Zhifeng Hao (Senior Member, IEEE) received the B.S. degree from Sun Yat-sen University, Guangzhou, China, and the Ph.D. degree in mathematics from Nanjing University, Nanjing, China, in 1990, and 1995, respectively. He is currently a Professor with the College of Science, Shantou University, Shantou, China. His research interests include algebra, machine learning, data mining, and evolutionary algorithms.

Lifang He (Member, IEEE) is currently an Assistant Professor with the Department of Computer Science and Engineering, Lehigh University, Bethlehem, PA, USA. Before her current position, he was a Postdoctoral Researcher with the Department of Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA. She has authored or coauthored more than 100 papers in refereed journals and conferences, such as NeurIPS, ICML, KDD, CVPR, WWW, IJCAI, AAAI, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON IMAGE PROCESSING, and AMIA. Her research interests include machine learning, data mining, tensor analysis, with major applications in biomedical data and neuroscience.