Integrating Vision-Language Semantic Graphs in Multi-View Clustering

JunLong Ke[1], **Zichen Wen**[1], **Yechenhao Yang**[1], **Chenhang Cui**[1],
Yazhou Ren[1,2*], **Xiaorong Pu**[1,2] and **Lifang He**[3]
[1]School of Computer Science and Engineering,University of Electronic Science and Technology of China
[2]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China
[3]Department of Computer Science and Engineering, Lehigh University
2021150901020@std.uestc.edu.cn, Zichen.Wen@outlook.com,
{yechenhaoyang, chenhangcui}@gmail.com,{yazhou.ren, puxiaor}@uestc.edu.cn, lih319@lehigh.edu

Abstract

In recent years, a variety of graph learning-based multi-view clustering (MVC) methods have emerged. However, these methods continue to face challenges in extracting latent features from real-world data, particularly in scenarios involving high-resolution color images and high-dimensional features. This task is notably difficult in cases where images are visually similar yet semantically diverse. To address this issue, we present a novel large-scale pre-trained model for multi-view clustering, named **I**ntegrate **V**ision-Language **S**emantic **G**raphs in **M**ulti-**V**iew Clustering (IVSGMV), which harnesses the capabilities of visual-language pre-training models to enhance clustering performance and confronts issues in the unsupervised tuning of pre-trained models for multi-view data. We introduce an effective unsupervised approach for creating semantic graphs from image multi-view datasets using pre-trained encoders. Our method addresses the inherent spatial noise and imbalance in these encoders by employing graph filters and a joint process that integrates both image node and edge features. Additionally, we demonstrate the application of our approach to multi-view image clustering on extensive datasets, notably the high-resolution MVImgNet, achieving an impressive 82% accuracy. Furthermore, our method extends the zero-shot capabilities of large-scale pre-trained models, resulting in good performance in clustering tasks on untrained multi-view datasets.

1 Introduction

In recent years, multi-view clustering (MVC) has emerged as a pivotal component in the realms of cross-modal representation learning and data-driven decision-making. This technique has been shown to be highly effective in various domains, such as image [Guan *et al.*, 2024b; Guan *et al.*, 2024a] and video analysis [Xu and Wei, 2021]. MVC methods capitalize on exploiting the consistency and complementary information of multiple views to group samples into distinct clusters [Ren *et al.*, 2022; Yan *et al.*, 2024; Wang *et al.*, 2021; Wang *et al.*, 2023; Zhang *et al.*, 2018a; Ling *et al.*, 2023]. Currently, self-supervised learning has achieved significant improvements in multi-view clustering [Zhou *et al.*, 2023; Zhou *et al.*, 2024b]. At the core of this approach lies the extraction and utilization of the intrinsic representational attributes of the data to further enhance clustering performance.

Nevertheless, previous MVC methods still encounter challenges when it comes to extracting latent features from real-world data, especially in scenarios involving high-resolution color images and high-dimensional features.

One potential resolution to these problems involves the incorporation of pre-trained models within the encoding process. Many approaches have used pre-trained models to explore the limits of clustering in single-view data [Adaloglou *et al.*, 2023; Shen *et al.*, 2023; Cai *et al.*, 2023]. However, there still exist challenges in the unsupervised tuning of pre-trained models for multi-view data. First, for k-means, it usually leads to unbalanced clustering [Yang *et al.*, 2017; Cui *et al.*, 2023] and is mainly applicable to data samples that are uniformly dispersed around the center [Van Gansbeke *et al.*, 2020]. On the other hand, when it comes to multi-view image data, that are similar in feature space do not always have the same semantic category [Van Gansbeke *et al.*, 2020] and therefore must be treated as noisy pairs, which will result in noisy spatial relations in images embedded.

In this paper, combining the classic image-text dual encoders model CLIP (Contrastive Language–Image Pretraining [Radford *et al.*, 2021]), we propose a novel multi-view clustering method using pre-trained models. In the face of the two previous limitations, we introduce the filter-based graph to utilize filters compatible with homogeneous and heterogeneous graphs to combat embedding noise and spatial imbalance, resulting in representations that balance spatially homophilous and heterophilous relationships. In addition, we use a joint graph to combine node features and edge relationships, where embeddings can be used to efficiently combine global feature relationships among multi-view data. To the best of our knowledge, our method is the first to apply large-scale pre-trained models multi-view data clustering.

Our main contributions can be summarized as follows:

- We introduce large-scale vision-language pre-trained models to multi-view clustering by proposing a method

*Corresponding author.

to counteract the embedded spatial noise and imbalance of the pre-trained encoder with graph filters and a graph joint process.

- Our method efficiently applies multi-view image clustering to large-scale multi-view image datasets, including the high-resolution multi-channel multi-view image dataset MVImgNet, achieving an accuracy of 82%.
- Our method achieves good performance for the clustering task on top of its untrained dataset, extending the zero-shot performance of CLIP.

2 Related Work

2.1 Multi-View Clustering

The collection of data from multiple perspectives and sources has become commonplace with the development of multimedia technology. Multi-view clustering (MVC) methods leverage complementary information from different views of the same instance to address the limitations of traditional clustering methods [Wu *et al.*, 2024; Ren *et al.*, 2024]. Several algorithms and techniques have been proposed for MVC, including classic methods like spectral clustering, as well as more recent developments in deep learning. Recently, deep learning has emerged as a powerful tool in the field of MVC, with numerous proposed architectures of deep neural networks tailored for this endeavor, including the introduction of new constraints, feature learning techniques, and graph filtering frameworks. LT-MSC [Zhang *et al.*, 2015] introduces a low-rank tensor constraint to explore the complementary information from multiple views. [Xu *et al.*, 2021b] proposes a framework for contrastive multi-view clustering (MFLVC) that utilizes multi-level feature learning to improve clustering effectiveness. [Wen *et al.*, 2024] suggests an adaptive hybrid graph filter based on homophily degree, which dynamically captures both low and high-frequency information to improve graph clustering. These methods have advanced the field of multi-view clustering significantly. However, there has been no exploration of incorporating the recently popular vision-language pre-training models into MVC, which has the potential to elevate multi-view clustering to unprecedented heights.

2.2 Vision-Language Pre-training Models

Vision-Language Pre-training (VLP) models aligning multi-modal data in a common feature space have been applied in various areas such as large vision language model [Zhou *et al.*, 2024a] and adversarial attack [Dong *et al.*, 2023]. They can be categorized into two main groups: those that use language-based training strategies, including Mask Language Modeling, such as mask language/region modeling VisualBert [Li *et al.*, 2019a], or autoregressive language modeling, such as image captioning and text-based image generation DALL-E [Ramesh *et al.*, 2021]. The other category is to utilize cross-modal contrastive learning to align the visual and textual information into a unified semantic space, e.g. CLIP [Radford *et al.*, 2021]. VLP aims to model the interaction between images and texts. Unicoder-VL [Li *et al.*, 2020] combines visual and textual embeddings, feeding them into a single encoder. In contrast, CLIP [Radford *et al.*, 2021] obtain visual and textual embeddings with separate encoders.

3 Method

As shown in Figures 1 and 2, the proposed method consists of two steps: 1) **Semantic Graph Construction**: Based on nodes characterized by images, we propose a method for constructing graph data, incorporating prior knowledge from the design of the CLIP model, we further construct edge relationship. This includes an unsupervised step of meaningful noun selection we refer to as *word filtering*, aiming to identify better nouns that accurately convey the overall details depicted in the image; 2) **Adaptive Hybrid Graph Filtering**: Herein, we design a filtering method that further processes based on the previously presented images and node-edge relationships. This method accounts for the heterogeneity arising from inconsistencies in their representations, resulting in an adaptive filter of Nodes and Edges that learns both consistency and complementarity, containing a graph joint process that combines features of nodes and edges, maximizes the use of consistency and complementarity across different views, while also leveraging that of node features and edge features as much as possible.

3.1 Semantic Graph Construction

Since the node feature of our graph to process is the embeddings of raw feature, it would be less meaningful when we directly use the similarity matrix between CLIP encoding results or original images as the adjacency matrix: the former does not take full advantage of the core of CLIP dual encoder design and training for computing image-text similarity, and the latter introduces noise in the edge relations because the original images are not encoded.

Given the application scenario of CLIP's design, a natural approach is to unsupervisedly select some words and then refer to the way CLIP predicts between them. By adding prefix words, we would calculate the cosine similarity between images and words to form the image-text bipartite graph.

In terms of word selection, to maximize the effectiveness of the final matrix representation, the meanings of the words should represent and distinguish the embeddings of the image dataset as much as possible without being overly generalized. In other words, this requires the identification of nouns that accurately convey the overall details depicted in the image. So first and foremost, we design a noun filter for selecting nouns unsupervisedly in the WordNet dataset [Miller, 1995] as raw noun set $\mathcal{N}_r$ to select a suitable noun set $\mathcal{N}$.

Noun Filter

Firstly, to exclude some nouns overly generalized, i.e., *Thing*, *Object*, *Item*, *Matter*, *Entity*, we initially exclude nouns from the WordNet that are close to the centers of almost all words. More precisely, we use cosine similarity between embeddings to measure the similarity of objects, as used in CLIP training. This selection can be formulated as selecting set $\mathcal{N}_d$ follows:

Let t be an element of the set $\mathcal{N}_d$. This membership is defined such that $t \in \mathcal{N}_d$ if and only if the Cosine Distance CD between the text encoder output $f(t, \theta_t)$ and the centroid $C_{\mathcal{N}}$ is greater than or equal to a threshold ε. Formally, we first select nouns t satisfied:

$$CD\left(f_t\left(t, \theta_t\right), C_{\mathcal{N}_r}\right) \geqslant \varepsilon, \tag{1}$$

we use set $\mathcal{N}_\mathbf{d}$ to mark the selected nouns t. Where, $f_t(\cdot)$ refers to the text encoder parameterized by θ_t, Cosine Distance between any two vectors $\mathbf{a}$ and $\mathbf{b}$ is defined as $CD(\mathbf{a}, \mathbf{b}) = 1 - \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$, ε refers to the threshold hyperparameter, in the following process and experimental setup, we maintain the threshold hyperparameter ε at 0.05. The centroid $C_\mathcal{N}$ is calculated as the mean of normalized text embeddings U_m for each noun m in the raw WordNet noun set N_r, expressed as:

$$C_{\mathcal{N}_r} = \frac{\sum_{m \in \mathcal{N}_r} \frac{U_m}{\|U_m\|}}{k}, \tag{2}$$

where k is the number of nouns in $\mathcal{N}_r$.

To further select an appropriate subset of nouns, considering other priors that can be integrated into the noun filter process that is merely the number of categories and the image embeddings themselves, it is noteworthy that experiments demonstrate over 70% accuracy in direct clustering of the embeddings of images from common datasets. Under these circumstances, for the given multi-view image datasets I subjected to clustering, we cluster the directly concatenated encoded results. Based on the clustering result, we select the final nouns $\mathcal{N}$ that satisfied:

$$CD\left(f_t\left(t, \theta_t\right), C_{Ii}\right) \leqslant \varepsilon_{th}, \tag{3}$$

where normalized centroids C_{Ii} are computed ba
embeddings of each dataset and their respective clu
bels as:

$$C_{Ii} = \frac{\sum_{CL_m = i, m \in I} \sum_{j=1}^{v} \frac{f_i(m_j, \theta_i)}{\|f_i(m_j, \theta_i)\|}}{v w_i},$$

where for any specific sample m, CL_m denotes th
signed to the sample by the k-means clustering
Here, w_i represents the number of samples in I wit
i, m_j refers to the j-th view of the sample m, an
number of views in I, $f_i(\cdot)$ refers to image encod
eterized by θ_i. Additionally, ε_{th} is defined as
threshold that is set such that the number of noun
the criterion outlined in Eq. (3) is equal to the hyper
β (i.e. the total number of nouns selected equal to
where N_{Class}refers to the classes number of imag
for each cluster label i.

In the following content, we call the selected nc
filtered nouns set $\mathcal{N}$.

Graph Construction with Filtered Nouns

Drawing upon the methodology of CLIP, particularly its approach to calculating result probabilities, we proceed to construct a comprehensive graph dataset utilizing the Filtered nouns obtained after the initial step. This is achieved by first creating a cosine similarity matrix derived from the embeddings of both the images and the Filtered nouns. Notably, diverging from the standard application of CLIP, in our approach, the constructed graph-text bipartite graph utilizes the cosine similarity matrix directly as the adjacency matrix $\mathbf{A}$, as opposed to the softmax operation typically employed in CLIP. This process is described by the following equation:

$$\begin{aligned} \mathbf{Z}_{I\,j}^{v} &= \mathbf{N}\left(f_i\left(m, \theta_i\right)\right), \\ \mathbf{Z}_{Nk} &= \mathbf{N}\left(f_t\left(t, \theta_t\right)\right), \\ \mathbf{B}^{v} &= \mathbf{Z}_I^{v} \mathbf{Z}_\mathcal{N}, \end{aligned} \tag{5}$$

where $\mathbf{B}^v$ represents the adjacency matrix of image-nouns bipartite graph based on cosine similarity in the v-th view. $\mathbf{Z}_{I\,j}^{v}$ refers to the j-th row of $\mathbf{Z}_I$ in v-th view, $\mathbf{Z}_{\mathcal{N}k}$ refers to the k-th row of $\mathbf{Z}_\mathcal{N}$, $m \epsilon I^v$, refer to the j-th image of dataset I 's v-th view, $t \epsilon I^v$, refer to the k-th nouns of filtered nouns set $\mathcal{N}$, $f_i(\cdot)$ refers to image encoder parameterized by θ_i, $f_t(\cdot)$ refers to text encoder parameterized by θ_t. $\mathbf{N}$ refers to row normalization.

Conclusively, within the computed image-text bipartite graph, since the image is our only clustering target, we have chosen an approach that involves multiplying the image-text similarity matrix with its transpose. This operation produces the final adjacency matrix $\mathbf{A}^v$ of v-th view, that is,

$$\mathbf{A}^v = \mathbf{B}^v \mathbf{B}^{vT}, \tag{6}$$

$\mathbf{A}^v$ would then be utilized for graph clustering. This strategy allows the clustering to also take into account the intrinsic connection between the image and its corresponding text.

For the above process, in the case of using CLIP, we use the pre-trained image encoder as $f_i(\cdot)$ and text encoder with the prompt template *A photo of a nouns* as $f_t(\cdot)$. We construct the adjacency matrix of the semantic graph to characterize edge relationships based on the multi-view dataset and unsupervised selection of nouns from WordNet.

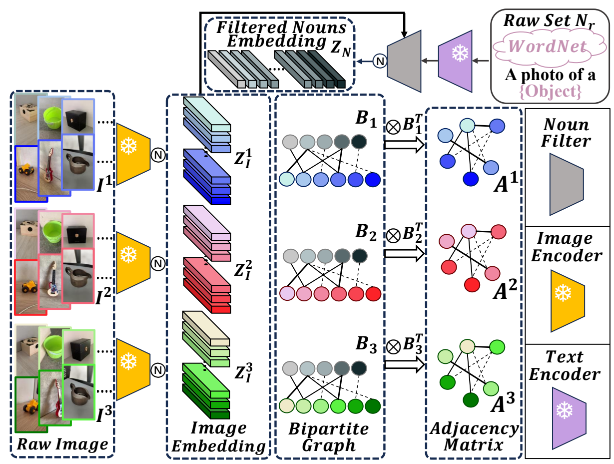


Figure 1: The overview of the graph construction process in IVS-GMV, showing an example of a 3-views dataset. For the v-th view, I^v refers to the raw image data, $\mathbf{Z}_I^v$ denotes the images' embedded features by image encoder, $\mathbf{Z}_n$ denotes the filtered nouns' embedded features by text encoder, and $\mathbf{B}^v$ is the adjacency matrix of image-nouns bipartite graph base on cosine similarity. All embeddings have been row normalized. The final adjacency matrix $\mathbf{A}^v$ is $\mathbf{B}^v$ and its transposed multiplication result.

3.2 Spectral Self-Supervised Learning with Adaptive Hybrid Graph Filter

Graph Joint Process

Considering that the semantic graph constructed in Section 3.1 may have some noisy edges, i.e., there may be inconsistencies between image embedding and word. It is feasible to use it directly for filtering, but its dominance in the filtering operation may lead to excessive loss of node feature information. Thus we try to utilize image embedding to further correct the semantic graph.

Specifically, we implement a graph joint process, which injects the nodes' original image features into the semantic graph. We first explore $\mathbf{A}$ from Eq. (6) as follows:

$$\mathbf{Z}_A = f_a(\mathbf{A}; \theta_a), \tag{7}$$

where $f_a(\cdot)$ represents deep auto-encoder, θ_a are learning parameters of the autoencoder. $\mathbf{Z}_A$ are the encoded outputs of $\mathbf{A}$. Since the graph data $\mathbf{A}^v$ of each view is a semantic graph generated by the same method, in order to motivate further extraction of inter-view consistency of the model, we adopt the practice of using a shared deep auto-encoder parameter θ_a for the semantic graphs of all views. We actually consider here the adjacency matrix $\mathbf{A}$ of the semantic graph being constructed as a kind of feature information, i.e., an edge feature, and thus we adopt a similar encoding operation as for image features.

Next, the image feature information is injected into the semantic graph. Consensus information between the image embedding and the edge features is then utilized to correct the semantic graph. To be more specific, we perform the following:

$$\begin{aligned} \mathbf{Z}_I &= f_i(I; \theta_i), \\ \mathbf{Z} &= \mathbf{Z}_A \mathbf{Z}_I^T, \\ \mathbf{S} = \mathbf{Z}\mathbf{Z}^T &= (\mathbf{Z}_A \mathbf{Z}_I^T)(\mathbf{Z}_A \mathbf{Z}_I^T)^T, \end{aligned} \tag{8}$$

where $f_i(\cdot)$ represents pre-trained image encoder with parameters θ_i, the dimension of $\mathbf{S}$ is the same as the original adjacency matrix $\mathbf{A}$ of semantic graph, we regard graph joint matrix $\mathbf{S}$ as the modified graph. $\mathbf{Z}_I$ is obtained through $f_{i\theta}$, $f_{i\theta}$ is the pre-trained image encoder. As $\mathbf{S}$ joins the image feature information of the nodes and the edge feature information of the semantic graph, it is more reliable than the adjacency matrix of the original semantic graph.

Additionally, we propose discretizing and sparsifying $\mathbf{S}$. This approach can save memory and accelerate computation, while also removing weakly correlated edge noise. We perform the following operation to remove the weakly correlated edge noise:

$$\begin{aligned} row\_means_i &= \frac{1}{N}\sum_{j=1}^{N} \mathbf{S}_{ij}, \\ \mathbf{S}_{\text{dis}_{ij}} &= \begin{cases} 1 & \text{if } \mathbf{S}_{ij} > row\_means_i, \\ 0 & \text{otherwise}, \end{cases} \end{aligned} \tag{9}$$

where N is the number of nodes and $\mathbf{S}_{\text{dis}}$ is the discretized $\mathbf{S}$. All $\mathbf{S}$ below denote the discretized $\mathbf{S}$, i.e., $\mathbf{S}_{\text{dis}}$ if not otherwise specified.

Adaptive Hybrid Graph Filter

After completing the graph joint process, a natural idea is to leverage popular graph neural networks [Zhou *et al.*, 2020] to explore structural information and feature information to enhance clustering performance. In terms of filter selection for graph filtering of the resulting semantic graph, the most common choice is to use the widely used Graph Convolutional Neural Network (GCN) to achieve this goal. Previous studies have pointed out that GCN is essentially a low-pass filter [Nt and Maehara, 2019]. From a spectral perspective, GCN only captures low-frequency information on the graph, completely losing high-frequency information. However, Bo *et al.* point out that both low-frequency and high-frequency information are equally crucial for learning representations of nodes on the graph [Bo *et al.*, 2021]. Completely discarding high-frequency information would lead to significant information loss, as in the constructed semantic graph, information is present both in the homogeneous and heterogeneous components.

A better idea is to use weighted high-pass and low-pass filters for obtaining high and low frequencies on the graph respectively to retain as much information as possible, and we design an adaptive hybrid graph filter as follows:

$$\begin{aligned} \widetilde{\mathbf{S}} = (\mathbf{D})^{-1}\mathbf{S}, \quad \widetilde{\mathbf{L}} = \mathbf{I} - \widetilde{\mathbf{S}}, \\ \mathbf{H}_{\mathbf{hybrid}} = hr \cdot (\widetilde{\mathbf{S}})^k \mathbf{Z}_I + (1 - hr) \cdot (\widetilde{\mathbf{L}})^k \mathbf{Z}_I, \end{aligned} \tag{10}$$

where $\mathbf{H}_{\mathbf{hybrid}}$ represents the output of the adaptive hybrid graph filter. hr is a learnable parameter that measures the homophily degree and is used to control the adaptive process of the hybrid graph filter, which will be calculated in Eq. (11). The diagonal matrix $\mathbf{D}_{ii} = \sum_j a_{ij}$ represents the degree matrix. $\widetilde{\mathbf{L}}$ is the normalized Laplace matrix. k is the order of the filter.

In Eq. (10), $(\widetilde{\mathbf{A}})^k \mathbf{Z}_I$ represents the low-pass filter and $(\widetilde{\mathbf{L}})^k \mathbf{Z}_I$ represents the high-pass filter.

Instead of manually setting the weights for the low-pass and high-pass filters, to further enhance the universality of the filtering mechanism, we have implemented an adaptive mechanism for the hybrid graph filter based on the homophily ratio. Homophily edges are edges on a graph that connect two similar nodes, and homophily ratio is a measure of the proportion of homophily edges on a graph. If a graph has a high homophily ratio, there will be more homophily edges, and from a graph signal processing perspective, low-frequency signals dominate the graph, and conversely, high-frequency signals dominate the graph. In other words, the homophily ratio of a graph can reflect the frequency distribution and the proportion of signals on the graph. Therefore, we considered assigning weights to hybrid graph filters using the homophily ratio and designing adaptive mechanisms. As the real label information is unavailable in the unsupervised setting, we estimate the homophily ratio (hr) using pseudo-labels:

$$hr = \frac{\text{SUM}(\mathbf{A}^v \odot \mathbf{P}\mathbf{P}^T - \mathbf{I})}{\text{SUM}(\mathbf{A}^v - \mathbf{I})}, \tag{11}$$

where, $\text{SUM}(\cdot)$ denotes the summation operation, $\odot$ denotes the Hadamard product, and $\mathbf{P} \in \{0,1\}^{n \times c}$ is the one-hot encoding of the pseudo label. Specifically, for the obtained semantic graph, if it has a high homophily ratio, then the low-pass filter in the adaptive hybrid filter will play a major role, and conversely, the high-pass filter will play a major role, which just matches the frequency distribution on the graph, thus extracting of information effectively. The final learned consensus embedding of n views is set as an adaptive weighted sum of the individual view embeddings after the graph filter: $\overline{\mathbf{H}} = \sum_{v=1}^{n} \omega_h^v \mathbf{H}^v$

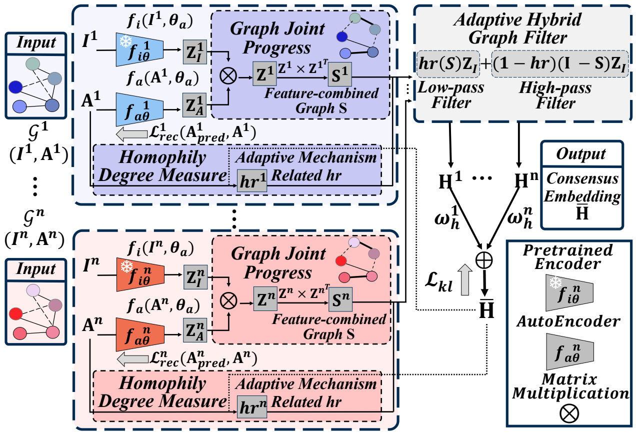


Figure 2: The illustration of the IVSGMV graph clustering framework. The inputs to the framework are the images I and the computed adjacency matrix $\mathbf{A}$ from the Semantic graph construction step. The final output of the framework is the consensus embedding $\overline{\mathbf{H}}$.

3.3 Model Optimization

In IVSGMV, learning from previous multi-view clustering studies [Zhao *et al.*, 2021; Ren *et al.*, 2022], in order to guide deep auto-encoder $f_a(\cdot)$ and learn consistency and complementarity between views, we apply reconstruction loss on $\mathbf{A}$ as follows:

$$\mathcal{L}_{Rec} = \mathcal{L}_{CE}(f_a(\mathbf{A};\theta_a);\mathbf{A}), \tag{12}$$

where $\mathcal{L}_{CE}(\cdot;\cdot)$ denotes cross-entropy loss.

We obtain the $\mathcal{L}_{KL}$ from the soft clustering distribution $\mathbf{Q}$ and the target distribution $\mathbf{P}$ as follows:

$$\mathcal{L}_{KL} = \sum_{v=1}^{V} KL(\overline{\mathbf{P}}\|\mathbf{Q}^v) + \sum_{v=1}^{V} KL(\mathbf{P}^v\|\mathbf{Q}^v) + KL(\overline{\mathbf{P}}\|\overline{\mathbf{Q}}), \tag{13}$$

where $\mathcal{L}_{KL}(\cdot;\cdot)$ denotes Kullback-Leibler (KL) divergence loss [Kullback and Leibler, 1951], $q_{ij}^v \in \mathbf{Q}^v$ describes the probability that node i in the v-th view belongs to the center of cluster j. $\mathbf{P}^v$ represents the target distribution of nodes embedding $\mathbf{H}^v$ in the v-th view. $\overline{\mathbf{Q}}$ and $\overline{\mathbf{P}}$ denote the soft and target distributions, respectively, of the consensus embedding $\overline{\mathbf{H}}$. $\mathcal{L}_{KL}$ encourages the soft distribution of each view to match the target distribution of the final consensus embedding $\overline{\mathbf{H}}$. Additionally, it enhances the consistency between the soft distribution and the target distribution of the consensus embedding.

Eventually, the loss of IVSGMV is defined as:

$$\mathcal{L} = \mathcal{L}_{Rec} + \mathcal{L}_{KL}, \tag{14}$$

4 Experiments

4.1 Datasets

As shown in Table 1, we use the following four real-world multi-view datasets in our study. MNIST [LeCun *et al.*, 1998] is a widely used dataset of handwritten digits from 0 to 9. The Fashion dataset [Xiao *et al.*, 2017] comprises images of various fashion items, including T-shirts, dresses, coats, etc. The COIL dataset [Nene *et al.*, 1996] contains images of various objects, such as cups, ducks, and blocks, shown in different poses. We use multi-view datasets derived from origin datasets: Multi-COIL-10, Multi-COIL-20, Multi-MNIST, and Multi-Fashion. Each dataset includes multiple views of each example, all randomly sampled from the same category. In Multi-COIL-10 (K = 10) and Multi-COIL-20 (K = 20), different views of an object correspond to various poses, but retain the same label. In Multi-MNIST, different views of a digit represent the same digit written by different individuals. In Multi-Fashion, different views of a product category signify different fashionable designs for that category. MVImgNet [Yu *et al.*, 2023] is a multi-view image dataset presented with a large scale, high accuracy, and large diversity, providing an average of 30 image views for each sample, based on its "MVlmgNet by categories" subset that is available in 2023, we select the most differentiated 3 views to build a 3-views images dataset in 14 categories.

Dataset	#Samples	#Views	#Clusters
Multi-MNIST	70000	2	10
Multi-Fashion	10000	3	10
Multi-COIL-10	720	3	10
Multi-COIL-20	1440	3	20
MVImgNet	24668	3	14

Table 1: The statistics of experimental datasets.

4.2 Comparison with State-of-the-Art Methods

Comparison Methods The comparison methods include three single-view clustering methods: K-means [MacQueen, 1967], β-VAE (β-VAE: learning basic visual concepts with a constrained variational framework [Higgins *et al.*, 2017]), and VaDE (variational deep embedding: an unsupervised and generative approach to clustering [Jiang *et al.*, 2017]), the input of which is the concatenation of all views, and five state-of-the-art MVC methods: BMVC (binary multi-view clustering [Zhang *et al.*, 2018b]), SAMVC (self-paced and auto-weighted multi-view clustering [Ren *et al.*, 2020]), RMSL (reciprocal multi-layer subspace learning for multi-view clustering [Li *et al.*, 2019b]), DEMVC (deep embedded multi-view clustering with collaborative training [Xu *et al.*, 2021a]), FMVACC (fast multi-view anchor-correspondence clustering [Wang *et al.*, 2022]), GCFAggMVC (Global and Cross-view Feature Aggregation for Multi-view Clustering [Yan *et al.*, 2023]), MFLVC (Multi-level feature learning for contrastive multi-view clustering [Xu *et al.*, 2022c]), DIMVC (Deep incomplete multi-view clustering via mining cluster complementarity [Xu *et al.*, 2022a]), SDMVC (Self-supervised discriminative feature learning for deep multi-view clustering [Xu *et al.*, 2022b])

Evaluation Metrics

We evaluate the effectiveness of clustering by three commonly used metrics, i.e., clustering accuracy (ACC), normalized mutual information (NMI), and adjusted rand index (ARI). A higher value of each evaluation metric indicates a better clustering performance.

Datasets	K-means	β-VAE	VaDE	BMVC	SAMVC	RMSL	DEMVC	FMVACC	**IVSGMV**
					ACC				
Multi-MNIST	53.9	49.3	30.7	89.3	-	-	<u>98.2</u>	55.6	**98.8**
Multi-Fashion	47.6	51.3	40.6	62.2	77.9	77.9	<u>78.6</u>	77.4	**93.3**
Multi-COIL-10	73.3	59.8	32.5	66.7	67.8	<u>96.4</u>	89.1	93.2	**100.0**
Multi-COIL-20	41.5	53.1	20.3	57.0	83.4	66.5	<u>85.0</u>	75.8	**99.9**
					NMI				
Multi-MNIST	48.2	43.6	35.4	90.2	-	-	**98.9**	48.2	<u>96.7</u>
Multi-Fashion	51.3	51.0	53.7	75.6	68.8	75.6	**90.3**	73.8	<u>89.0</u>
Multi-COIL-10	76.9	68.5	44.8	68.1	82.6	92.5	89.7	93.4	**100.0**
Multi-COIL-20	64.5	66.7	36.9	90.0	79.1	76.3	<u>96.5</u>	84.9	**99.9**
					ARI				
Multi-MNIST	36.0	29.1	8.5	85.6	-	-	<u>98.6</u>	45.2	**98.8**
Multi-Fashion	34.8	33.7	22.8	68.2	55.7	68.2	<u>77.2</u>	70.3	**87.1**
Multi-COIL-10	64.8	51.4	18.1	53.0	62.1	92.1	89.7	<u>92.5</u>	**100.0**
Multi-COIL-20	38.4	45.0	9.0	81.3	55.4	58.7	<u>86.0</u>	79.1	**99.9**

Table 2: Clustering results of methods on four common datasets. The best result in each row is shown in **bold** and the second-best is <u>underlined</u>.

Datasets	K-means	GCFAggMVC	MFLVC	DIMVC	SDMVC	**IVSGMV**
			ACC			
MVImgNet(VIT/L)	<u>73.6</u>	36.1	34.5	68.9	69.7	**82.1**
MVImgNet(DS)	20.2	24.8	22.3	14.4	25.4	
			NMI			
MVImgNet(VIT/L)	77.3	53.3	63.2	78.1	<u>81.5</u>	**81.8**
MVImgNet(DS)	12.7	15.5	13.4	5.08	16.0	
			ARI			
MVImgNet(VIT/L)	64.8	**83.2**	34.0	65.9	67.4	<u>75.6</u>
MVImgNet(DS)	5.39	8.07	4.23	1.40	7.55	

Table 3: Clustering results of various methods on MVImgNet dataset. Where MVImgNet (VIT/L) refers to using the image embeddings of CLIP ViT-L/14@336px model encoder as input, MVImgNet (DS) refers to using the downsampling of the dataset at the highest possible resolution under the same 48GB memory limit as input. The best results for each row are shown in **bold** and the second-best results are <u>underlined</u>.

Results

Tables 2 and 3 shows the quantitative comparison between the proposed method and baseline models for several datasets. IVSGMV achieved superior performance compared to baselines on all datasets, and the state-of-the-art comparison method still mighty underperforms our proposed method on the multichannel high-resolution multi-view dataset MVImgNet even when CLIP encoding embeddings are used as inputs. This reflects our full application of large-scale pre-trained model CLIP.

Compenents / Datasets	Multi-Fashion ACC	NMI	ARI
IVSGMV (w/o $\mathbf{A}$ w/ $\mathbf{A}_e$)	91.8	87.5	85.1
IVSGMV (w/o $\mathbf{A}$ w/ $\mathbf{A}_r$)	89.5	86.7	83.4
IVSGMV (w/o $f_{i\theta}$)	83.5	79.7	73.7
IVSGMV (w/o $f_{i\theta}$&$\mathbf{A}$ w/ $\mathbf{A}_e$)	82.2	77.6	71.6
IVSGMV (w/o $f_{i\theta}$&$\mathbf{A}$ w/ $\mathbf{A}_r$)	76.2	70.1	65.3
IVSGMV (w/o $\mathbf{S}$)	90.8	85.5	82.1
IVSGMV (w/o $\mathbf{H}_{\mathbf{hybrid}}$)	92.5	88.2	85.6
IVSGMV (w/o $\mathbf{S}$&$\mathbf{H}_{\mathbf{hybrid}}$)	89.8	84.6	80.9
IVSGMV	**93.3**	**89.0**	**87.1**

Table 4: The ablation study results of IVSGMV on Multi-Fashion. The original results are shown in **bold**.

Vision-Language Model Implementation

For the pre-trained encoder, we used the CLIP ViT-L/14@336px model, whose visual and text backbones are ViT [Dosovitskiy *et al.*, 2020].

4.3 Ablation Studies

Effect of Semantic Graph Construction and Pre-trained Encoder

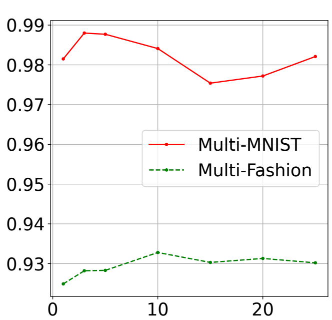
To understand the importance of our graph construction method and pre-trained encoder. We conducted three ablation experiments on both to analyze their impact on the performance of the IVSGMV in the Multi-Fashion dataset, where $f_{i\theta}$ represents the use of pre-trained self-encoder as node feature and $w/of_{i\theta}$ represents the direct use of the original image features without encoder, $\mathbf{A}$ represents the use of the proposed adjacency matrix construction method. As alternative constructions of $\mathbf{A}$, $\mathbf{A}_e$ and $\mathbf{A}_r$ refer to graph construction method accord to the spectral clustering, representing the use of affinity matrix based on cosine similarity constructed using the raw features ($\mathbf{A}_r$) and image embeddings ($\mathbf{A}_e$) of the pre-trained encoder, respectively.

Effect of Graph Process Components

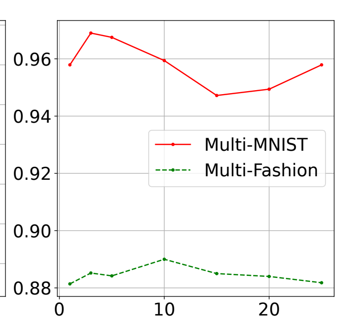
Graph joint aggregation matrix $\mathbf{S}$ and adaptive hybrid graph filter are important components of the IVSGMV. We conducted three more ablation experiments. w/o $\mathbf{S}$ represents the alternative implementation of graph joint aggregation matrix $\mathbf{S}$ with adjacency matrix $\mathbf{A}$, w/o $\mathbf{H_{hybrid}}$ represents the alternative implementation of adaptive hybrid graph filter with a common GCN low-pass filter $\mathbf{H}_{lp}$ and w/o $\mathbf{S}$&$\mathbf{H_{hybrid}}$ represents the combination of the above two. As Table 4 demonstrates, the performance of the model is greatly and adversely affected whether the graph joint aggregation matrix or adaptive hybrid graph filter, or both are eliminated.

The results of ablation Table 4 show that compared to $\mathbf{A}_e$ and $\mathbf{A}_r$, the semantic graph construction method of $\mathbf{A}$ leads to 1.5% and 3.8% ACC improvement, and the pre-trained encoder $f_{i\theta}$ brings about 18.7% accuracy improvement. Also, the graph joint aggregation and hybrid graph filter lead to 0.8% and 2.5% ACC improvement.

Notably, the semantic graph construction method of $\mathbf{A}$ also delivers 1.3% and 7.3% ACC improvement compared to $\mathbf{A}_e$ and $\mathbf{A}_r$ without the use of $f_{i\theta}$, respectively, which demonstrates the methodological superiority of semantic graph construction.



(a) ACC



(b) NMI

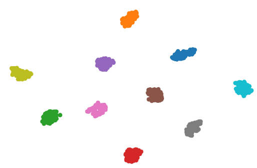
Figure 3: Parameter sensitivity analysis about *order* on Multi-MNIST and Multi-Fashion.

Parameter Sensitivity Analysis

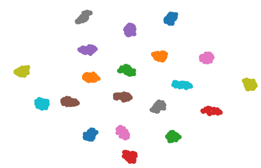
The sensitivity analysis for *order* is on Figure 3. From the spatial perspective, *order* controls the aggregation order of the graph filter. The higher *order* enables nodes to aggregate information from more distant ones, while nodes can only access feature information of closer nodes in lower *order*.

Visualization of Consensus Embedding $\overline{\mathbf{H}}$

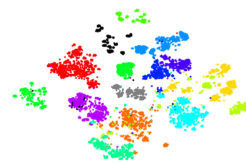
Figure 4 visualizes the consensus embedding $\overline{\mathbf{H}}$ of our model on Multi-COIL-10, Multi-COIL-20, and MVImgNet.



(a) Multi-COIL-10



(b) Multi-COIL-20



(c) MVImgNet

Figure 4: Visualization of learned consensus embedding $\overline{\mathbf{H}}$ on Multi-COIL-10, Multi-COIL-20, and MVImgNet.

5 Conclusion

In this study, we introduce large-scale pre-trained models into the field of multi-view clustering. We analyze the challenges of unsupervised fine-tuning pre-trained models and combat imbalance and noise relations in the embedding space with the help of building semantic graphs and graph filters. For semantic graph building, we propose an unsupervised construction process based on the CLIP model priori, which builds a graph-word bipartite graph by adaptively selecting words from the WordNet, and further obtains the adjacency matrices that imply semantic relations. On the graph clustering model, we apply an adaptive hybrid graph filter for multi-view clustering to adaptively mine the low and high-frequency information in the graph to learn distinguishable node embeddings, which in turn mines the homogeneous and heterogeneous information among embedding relations of the pre-trained model. In addition, the joint graph process is used to construct filters to enhance the distinguishability between low and high-frequency signals, where threshold discretization is applied to combat noise. Our IVSGMV has good performance on several multi-view datasets, extending the application of multi-view clustering to realistic image datasets while maintaining competitive performance on other datasets.

Acknowledgements

This work is supported in part by Shenzhen Science and Technology Program under grants JCYJ20230807115959041 and JCYJ20230807120010021. Lifang He is partially supported by the NSF grants (MRI-2215789, IIS-1909879, IIS-2319451), NIH grant under R21EY034179, and Lehigh's grants under Accelerator and CORE.

References

[Adaloglou *et al.*, 2023] Nikolas Adaloglou, Felix Michels, Hamza Kalisch, and Markus Kollmann. Exploring the limits of deep image clustering using pretrained models. *arXiv preprint arXiv:2303.17896*, 2023.

[Bo *et al.*, 2021] Deyu Bo, Xiao Wang, Chuan Shi, and Huawei Shen. Beyond low-frequency information in graph convolutional networks. In *AAAI*, pages 3950–3957, 2021.

[Cai *et al.*, 2023] Shaotian Cai, Liping Qiu, Xiaojun Chen, Qin Zhang, and Longteng Chen. Semantic-enhanced image clustering. In *AAAI*, pages 6869–6878, 2023.

[Cui *et al.*, 2023] Chenhang Cui, Yazhou Ren, Jingyu Pu, Xiaorong Pu, and Lifang He. Deep multi-view subspace clustering with anchor graph. In *IJCAI*, pages 3577–3585, 2023.

[Dong *et al.*, 2023] Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How robust is google's bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*, 2023.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Guan *et al.*, 2024a] Renxiang Guan, Zihao Li, Xianju Li, and Chang Tang. Pixel-superpixel contrastive learning and pseudo-label correction for hyperspectral image clustering. In *ICASSP*, pages 6795–6799, 2024.

[Guan *et al.*, 2024b] Renxiang Guan, Zihao Li, Wenxuan Tu, Jun Wang, Yue Liu, Xianju Li, Chang Tang, and Ruyi Feng. Contrastive multi-view subspace clustering of hyperspectral images based on graph convolutional networks. *TGRS*, 62:1–14, 2024.

[Higgins *et al.*, 2017] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

[Jiang *et al.*, 2017] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: an unsupervised and generative approach to clustering. In *IJCAI*, pages 1965–1972, 2017.

[Kullback and Leibler, 1951] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324, 1998.

[Li *et al.*, 2019a] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[Li *et al.*, 2019b] Ruihuang Li, Changqing Zhang, Huazhu Fu, Xi Peng, Tianyi Zhou, and Qinghua Hu. Reciprocal multi-layer subspace learning for multi-view clustering. In *ICCV*, pages 8172–8180, 2019.

[Li *et al.*, 2020] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020.

[Ling *et al.*, 2023] Yawen Ling, Jianpeng Chen, Yazhou Ren, Xiaorong Pu, Jie Xu, Xiaofeng Zhu, and Lifang He. Dual label-guided graph refinement for multi-view graph clustering. In *AAAI*, pages 8791–8798, 2023.

[MacQueen, 1967] James MacQueen. Classification and analysis of multivariate observations. In *BSMSP*, pages 281–297, 1967.

[Miller, 1995] George A Miller. Wordnet: a lexical database for english. *CACM*, 38(11):39–41, 1995.

[Nene *et al.*, 1996] Sameer A Nene, Shree K Nayar, Hiroshi Murase, et al. Columbia object image library (coil-20). *Technical Report, CUCS-005-96*, 1996.

[Nt and Maehara, 2019] Hoang Nt and Takanori Maehara. Revisiting graph neural networks: All we have is low-pass filters. *arXiv preprint arXiv:1905.09550*, 2019.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[Ramesh *et al.*, 2021] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831, 2021.

[Ren *et al.*, 2020] Yazhou Ren, Shudong Huang, Peng Zhao, Minghao Han, and Zenglin Xu. Self-paced and auto-weighted multi-view clustering. *Neurocomputing*, 383:248–256, 2020.

[Ren *et al.*, 2022] Yazhou Ren, Jingyu Pu, Zhimeng Yang, Jie Xu, Guofeng Li, Xiaorong Pu, Philip S Yu, and Lifang He. Deep clustering: A comprehensive survey. *arXiv preprint arXiv:2210.04142*, 2022.

[Ren *et al.*, 2024] Yazhou Ren, Xinyue Chen, Jie Xu, Jingyu Pu, Yonghao Huang, Xiaorong Pu, Ce Zhu, Xiaofeng Zhu, Zhifeng Hao, and Lifang He. A novel federated multi-view clustering method for unaligned and incomplete data fusion. *Information Fusion*, 108:102357, 2024.

[Shen *et al.*, 2023] Shuai Shen, Wanhua Li, Xiaobing Wang, Dafeng Zhang, Zhezhu Jin, Jie Zhou, and Jiwen Lu. Clip-cluster: Clip-guided attribute hallucination for face clustering. In *CVPR*, pages 20786–20795, 2023.

[Van Gansbeke *et al.*, 2020] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *ECCV*, pages 268–285, 2020.

[Wang *et al.*, 2021] Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. Generative partial multi-view clustering with adaptive fusion and cycle consistency. *TIP*, 30:1771–1783, 2021.

[Wang *et al.*, 2022] Siwei Wang, Xinwang Liu, Suyuan Liu, Jiaqi Jin, Wenxuan Tu, Xinzhong Zhu, and En Zhu. Align then fusion: Generalized large-scale multi-view clustering with anchor matching correspondences. In *NeurIPS*, pages 5882–5895, 2022.

[Wang *et al.*, 2023] Qianqian Wang, Zhiqiang Tao, Wei Xia, Quanxue Gao, Xiaochun Cao, and Licheng Jiao. Adversarial multiview clustering networks with adaptive fusion. *TNNLS*, 34:7635–7647, 2023.

[Wen *et al.*, 2024] Zichen Wen, Yawen Ling, Yazhou Ren, Tianyi Wu, Jianpeng Chen, Xiaorong Pu, Zhifeng Hao, and Lifang He. Homophily-related: Adaptive hybrid graph filter for multi-view graph clustering. In *AAAI*, pages 15841–15849, 2024.

[Wu *et al.*, 2024] Song Wu, Yan Zheng, Yazhou Ren, Jing He, Xiaorong Pu, Shudong Huang, Zhifeng Hao, and Lifang He. Self-weighted contrastive fusion for deep multi-view clustering. *TMM*, 2024.

[Xiao *et al.*, 2017] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[Xu and Wei, 2021] Yuanjin Xu and Ming Wei. Multi-view clustering toward aerial images by combining spectral analysis and local refinement. *Future Generation Computer Systems*, 117:138–144, 2021.

[Xu *et al.*, 2021a] Jie Xu, Yazhou Ren, Guofeng Li, Lili Pan, Ce Zhu, and Zenglin Xu. Deep embedded multi-view clustering with collaborative training. *Information Sciences*, 573:279–290, 2021.

[Xu *et al.*, 2021b] Jie Xu, Yazhou Ren, Huayi Tang, Xiaorong Pu, Xiaofeng Zhu, Ming Zeng, and Lifang He. Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering. In *ICCV*, pages 9234–9243, 2021.

[Xu *et al.*, 2022a] Jie Xu, Chao Li, Yazhou Ren, Liang Peng, Yujie Mo, Xiaoshuang Shi, and Xiaofeng Zhu. Deep incomplete multi-view clustering via mining cluster complementarity. In *AAAI*, pages 8761–8769, 2022.

[Xu *et al.*, 2022b] Jie Xu, Yazhou Ren, Huayi Tang, Zhimeng Yang, Lili Pan, Yang Yang, Xiaorong Pu, Philip S Yu, and Lifang He. Self-supervised discriminative feature learning for deep multi-view clustering. *TKDE*, 2022.

[Xu *et al.*, 2022c] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *CVPR*, pages 16051–16060, 2022.

[Yan *et al.*, 2023] Weiqing Yan, Yuanyang Zhang, Chenlei Lv, Chang Tang, Guanghui Yue, Liang Liao, and Weisi Lin. Gcfagg: Global and cross-view feature aggregation for multi-view clustering. In *CVPR*, pages 19863–19872, 2023.

[Yan *et al.*, 2024] Wenbiao Yan, Yiyang Zhou, Yifei Wang, Qinghai Zheng, and Jihua Zhu. Multi-view semantic consistency based information bottleneck for clustering. *Knowledge-Based Systems*, 288:111448, 2024.

[Yang *et al.*, 2017] Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *ICML*, pages 3861–3870, 2017.

[Yu *et al.*, 2023] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *CVPR*, pages 9150–9161, 2023.

[Zhang *et al.*, 2015] Changqing Zhang, Huazhu Fu, Si Liu, Guangcan Liu, and Xiaochun Cao. Low-rank tensor constrained multiview subspace clustering. In *ICCV*, pages 1582–1590, 2015.

[Zhang *et al.*, 2018a] Changqing Zhang, Huazhu Fu, Qinghua Hu, Xiaochun Cao, Yuan Xie, Dacheng Tao, and Dong Xu. Generalized latent multi-view subspace clustering. *TPAMI*, 42(1):86–99, 2018.

[Zhang *et al.*, 2018b] Zheng Zhang, Li Liu, Fumin Shen, Heng Tao Shen, and Ling Shao. Binary multi-view clustering. *TPAMI*, 41:1774–1782, 2018.

[Zhao *et al.*, 2021] Han Zhao, Xu Yang, Zhenru Wang, Erkun Yang, and Cheng Deng. Graph debiased contrastive learning with joint representation clustering. In *IJCAI*, pages 3434–3440, 2021.

[Zhou *et al.*, 2020] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020.

[Zhou *et al.*, 2023] Yiyang Zhou, Qinghai Zheng, Shunshun Bai, and Jihua Zhu. Semantically consistent multi-view representation learning. *Knowledge-Based Systems*, 278:110899, 2023.

[Zhou *et al.*, 2024a] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *ICLR*, 2024.

[Zhou *et al.*, 2024b] Yiyang Zhou, Qinghai Zheng, Yifei Wang, Wenbiao Yan, Pengcheng Shi, and Jihua Zhu. Mcoco: Multi-level consistency collaborative multi-view clustering. *Expert Systems with Applications*, 2024.