Incomplete Multi-view Clustering via Diffusion Contrastive Generation

Yuanyang Zhang[1*], Yijie Lin[2*], Weiqing Yan[3], Li Yao[1,4†], Xinhang Wan[5],
Guangyuan Li[6], Chao Zhang[7], Guanzhou Ke[8], Jie Xu[9]
[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]College of Computer Science, Sichuan University
[3]School of Computer and Control Engineering, Yantai University
[4]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
[5]College of Computer, National University of Defense Technology
[6]College of Computer Science and Technology, Zhejiang University
[7]Department of Control Science and Intelligence Engineering, Nanjing University
[8]School of Economics and Management, Beijing Jiaotong University
[9]School of Computer Science and Engineering, University of Electronic Science and Technology of China
{zhangyuanyang, yao.li}@seu.edu.cn, linyijie.gm@gmail.com

Abstract

Incomplete multi-view clustering (IMVC) has garnered increasing attention in recent years due to the common issue of missing data in multi-view datasets. The primary approach to address this challenge involves recovering the missing views before applying conventional multi-view clustering methods. Although imputation-based IMVC methods have achieved significant improvements, they still encounter notable limitations: 1) heavy reliance on paired data for training the data recovery module, which is impractical in real scenarios with high missing data rates; 2) the generated data often lacks diversity and discriminability, resulting in suboptimal clustering results. To address these shortcomings, we propose a novel IMVC method called Diffusion Contrastive Generation (DCG). Motivated by the consistency between the diffusion and clustering processes, DCG learns the distribution characteristics to enhance clustering by applying forward diffusion and reverse denoising processes to intra-view data. By performing contrastive learning on a limited set of paired multi-view samples, DCG can align the generated views with the real views, facilitating accurate recovery of views across arbitrary missing view scenarios. Additionally, DCG integrates instance-level and category-level interactive learning to exploit the consistent and complementary information available in multi-view data, achieving robust and end-to-end clustering. Extensive experiments demonstrate that our method outperforms state-of-the-art approaches.

Introduction

In practical applications, data collected from different sensors or feature extraction methods is often presented in multiple-view forms. For instance, healthcare professionals utilize multi-view data such as MRI images, CT scans, and X-rays for diagnostic purposes (Chen 2022). Similarly, autonomous vehicles employ data from cameras, LiDAR, and radar to perceive their surroundings and navigate (Cui et al. 2024). Multi-view data offers a comprehensive perspective of the same subject from various angles, making it a critical area of machine learning and data mining.

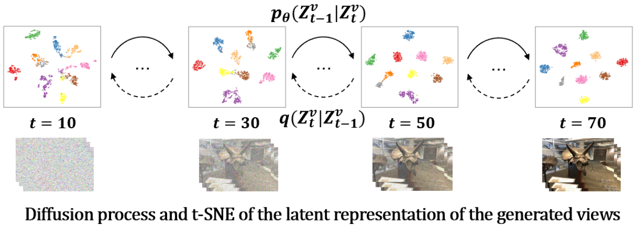

$p_\theta(Z_{t-1}^v|Z_t^v)$
…
…
…
$q(Z_t^v|Z_{t-1}^v)$
$t = 10$
$t = 30$
$t = 50$
$t = 70$
Diffusion process and t-SNE of the latent representation of the generated views

Figure 1: Motivation. We observed that as the diffusion process progresses, the generated sample representations tend to gradually aggregate toward the center of the cluster.

To extract and synthesize the full picture from these disparate sources, one of the pivotal techniques in multi-view learning is multi-view clustering (MVC), which aims to partition multi-view data into distinct clusters (Cai et al. 2024a; Yan et al. 2024; Cai et al. 2024b; Lu et al. 2024) by leveraging the inherent consistency and complementarity nature of the information across different views (Lin et al. 2024). The success of existing multi-view clustering methods heavily relies on the completeness of multi-view data, i.e., all views are consistently available for each sample. However, this condition is rarely met in real-life scenarios due to sensor malfunctions or damage during storage. To tackle this problem, Incomplete Multi-view Clustering (IMVC) has been proposed to effectively cluster data even when some views are missing (Zhang et al. 2023; Wan et al. 2024).

The predominant strategy for addressing the incomplete multi-view data is to recover the missing view and subsequently apply existing MVC methods (Guan et al. 2024; Ke et al. 2024; Sun et al. 2024; Cai et al. 2022) for clustering. Specifically, existing IMVC methods primarily focus on crafting consistent common representations through con-

[*] These authors contributed equally.
[†] Corresponding authors

trastive learning and various data generation techniques to impute and fill in the missing data. Among these, researchers have explored several imputation-based approaches, including cross-view neighbors-based methods (Tang and Liu 2022; Jin et al. 2023), cross-view prediction-based methods (Lin et al. 2021, 2022), and generative adversarial network-based methods (Wang et al. 2021). Despite significant improvements achieved by these methods, they still encounter the following issues: 1) heavy reliance on paired data for training is difficult to satisfy in practical scenarios, especially in cases with high missing rates. For instance, meteorological monitoring stations are usually located in remote areas, making them vulnerable to weather and power supply issues, which results in high amounts of missing data; 2) the generated data often lacks diversity and discriminability, leading to suboptimal clustering results. For example, data recovery based on neighbor prediction may lack diversity, while that based on adversarial generation may lack discriminability due to the instability of model training.

Without necessitating adversarial training, diffusion models are renowned for their robust generative capabilities, presenting a promising alternative. For example, Wen et al. (2024) first incorporates the diffusion model into IMVC by using the available views as the condition for diffusion to generate the missing views. However, this approach still heavily relies on paired data for diffusion completion and ignores the necessary connections between clustering and data completion.

To alleviate the above problems, we propose a novel incomplete multi-view clustering method termed diffusion contrastive generation (DCG), inspired by the consistency between the inverse diffusion process in data generation and the compactness of sample clusters. Our motivation is shown in Figure 1, as the diffusion process progresses, the generated sample representations gradually converge to the center of the category to which they belong. This phenomenon reveals the implicit connection between the diffusion process and data clustering, i.e., the inverse diffusion process not only facilitates data generation but also enhances the compactness of the data clusters.

Based on this phenomenon, DCG first applies forward diffusion and reverse denoising to the intra-view data to learn the distribution characteristics for learning clustering-friendly representations. Then, by performing contrastive learning on only a small number of paired samples, we can align the generated views with the real views, enabling view recovery in arbitrary view missing scenarios. During the inference stage, by simply extrapolating the diffusion step (e.g., from 50 to 100), DCG obtains more desirable clustering results thanks to the diffusion and clustering consistency. Through DCG, we not only achieve view completion with only a small amount of paired data but also integrate the data completion and clustering processes into a unified diffusion process. Additionally, DCG integrates instance-level and category-level interactive learning to fully utilize consistent and complementary information in multi-view data, achieving end-to-end clustering. The main contributions of this paper are summarized as follows:

- We reveal that the diffusion process not only facilitates the recovery of missing views but also enhances the compactness of data clustering as the diffusion progresses. By extrapolating the diffusion step during the inference stage, we can obtain more compact clustering results thanks to the consistency between the diffusion process and clustering.
- The proposed DCG method innovatively combines the diffusion process with contrastive learning to enable effective view generation and discriminative learning with limited paired data.
- DCG achieves significant improvements across multiple datasets, especially in cases of high missing rates, where it outperforms state-of-the-art methods by 6.67% on the CUB dataset with the missing rate of 0.7.

Related Work

In this section, we briefly review recent advancements in two related topics, namely, deep incomplete multi-view clustering and diffusion model.

Deep Incomplete Multi-view Clustering

Inspired by the powerful feature representation capabilities of deep learning, many DIMVC methods have been developed. These methods can be categorized into two types: imputation-free methods and imputation-based methods. (1) Imputation-free methods aggregate the representations of existing views through different approaches and then obtain clustering results from the aggregated representations (Xu et al. 2023, 2024a). (2) Imputation-based methods employ various strategies to impute missing views and subsequently perform clustering on the completed multi-view dataset. Specific methods include: i) cross-view neighbor-based approaches (Tang and Liu 2022; Jin et al. 2023) that use cross-view nearest methods to impute missing data; ii) cross-view prediction-based methods (Lin et al. 2021, 2022) that maximize mutual information between different views of the samples and uses an encoder to estimate the missing views; iii) generative adversarial network-based methods (Wang et al. 2021) that recover missing data through generation adversarial network.

Diffusion Model

Diffusion Model (Ho, Jain, and Abbeel 2020) is a generative model that has achieved significant success in fields such as image generation (Li et al. 2024b; Xu et al. 2024b), image inpainting (Lugmayr et al. 2022), and image super-resolution (Li et al. 2024a). It utilizes a Markov diffusion process, gradually injecting Gaussian noise into the data, and then generates samples through a reverse denoising process. This approach has yielded impressive results in generating high-quality images. For instance, Li et al. utilize diffusion models to generate accurate prior knowledge, guiding the reconstruction of MRI images to maintain high quality without distortion (Li et al. 2024a). Generative Adversarial Networks (GANs) have achieved success in incomplete multi-view clustering (Wang et al. 2021) but face challenges such as training instability and model collapse (Liu

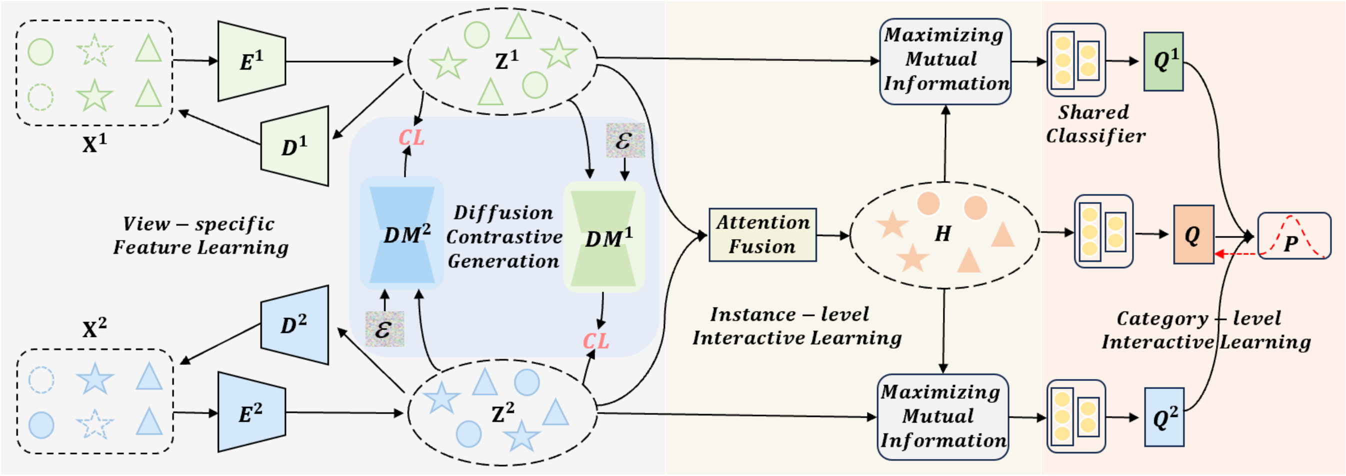

X¹
E¹
D¹
Z¹
CL
ℰ
DM²
Diffusion Contrastive Generation
DM¹
View − specific Feature Learning
ℰ
CL
X²
D²
E²
Z²
Maximizing Mutual Information
Shared Classifier
Q¹
Attention Fusion
H
Q
P
Instance − level Interactive Learning
Category − level Interactive Learning
Maximizing Mutual Information
Q²

Figure 2: Without loss of generality, we take bi-view data as a showcase to demonstrate the overall framework of our proposed DCG. As shown, our method is mainly divided into four modules: (1) View-specific Feature Learning; (2) Diffusion Contrastive Generation; (3) Instance-level Interactive Learning; (4) Category-level Interactive Learning. Note that, *E*: Encoder; *D*: Decoder; *DM*: Diffusion Model; *CL*: Contrastive Learning.

et al. 2023; Luo and Yang 2024). In contrast, diffusion models based on Markov diffusion processes offer advantages of stable training processes and diverse generated results. Inspired by these studies, we propose the DCG approach, which aims to utilize the powerful generative capabilities of diffusion models to recover missing views from existing view information.

Method

Notations. Given a multi-view incomplete dataset $\{\mathbf{X}^v \in \mathbb{R}^{N_v \times D_v}\}_{v=1}^V$ which consists of N_v samples across V views, where $\mathbf{X}^v = \{x_1^v, x_2^v, \ldots, x_{N_v}^v\} \in \mathbb{R}^{N_v \times D_v}$ denote the instance set of the v-th view, D_v is the dimensionality of the samples, and N_v denotes the number of samples. K is the number of clusters.

View-specific Feature Learning

Deep autoencoder is widely used for unsupervised representation learning (Lin et al. 2021) by minimizing the reconstruction error. Considering that different views contain specific feature information, we construct a view-specific autoencoder to learn the latent representation $\mathbf{Z}^v \in \mathbb{R}^{N_v \times d_v}$ by minimizing the reconstruction loss:

$$\mathcal{L}_{\mathbf{R}} = \sum_{v=1}^{V}\sum_{i=1}^{N}\left\|\mathbf{X}_i^v - D_{\phi^v}^{(v)}\left(\mathbf{Z}_i^v\right)\right\|_2^2, \tag{1}$$

where $\mathbf{X}_i^v$ denotes the i-th sample of $\mathbf{X}^v$, and $D_{\phi^v}^{(v)}$ is the decoder of v-th view with the parameters ϕ^v. The latent representation $\mathbf{Z}_i^v$ is obtained as follows:

$$\mathbf{Z}_i^v = E_{\eta^v}^{(v)}\left(\mathbf{X}_i^v\right). \tag{2}$$

where $E^{(v)}$ is the encoder of v-th view with parameters η^v.

Diffusion Contrastive Generation

To generate accurate missing view representations, inspired by the diffusion model (Ho, Jain, and Abbeel 2020) and contrastive learning (Chen et al. 2020; Lin et al. 2023), we designed a Diffusion Contrastive Generation module. Specifically, the module first applies forward diffusion and reverse denoising on the intra-view data to learn the distribution characteristics for clustering. Subsequently, by performing contrastive learning on a small number of paired samples, we can align the generated views with the real views, enabling view recovery in arbitrary view missing scenarios. The Diffusion Contrastive Generation module includes the forward diffusion process, the reverse denoising process, and the data recovery process.

Forward Diffusion Process. During the forward diffusion process, we gradually add Gaussian noise to the initial latent representation $\mathbf{Z}_0^v$ until it becomes fully Gaussian noise:

$$q\left(\mathbf{Z}_T^v \mid \mathbf{Z}_0^v\right) = \mathcal{N}\left(\mathbf{Z}_T^v; \sqrt{\bar{\alpha}_T}\mathbf{Z}_0^v, \left(1-\bar{\alpha}_T\right)\mathbf{I}\right), \tag{3}$$

where T is the total number of time steps, $\mathcal{N}$ denotes the Gaussian distribution, α and $\bar{\alpha}_T$ are defined as: $\alpha = 1 - \beta_t, \quad \bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$, where $t = (1, \ldots, T)$, $\beta_{1:T} \in (0, 1)$ are hyperparameters that control the variance of the noise.

Reverse Denoising Process. The reverse denoising process is a Markov chain running backwards from $\mathbf{Z}_T^v$ to $\hat{\mathbf{Z}}^v$. Taking the reverse step from $\mathbf{Z}_t^v$ to $\mathbf{Z}_{t-1}^v$ as an example:

$$p_\theta\left(\mathbf{Z}_{t-1}^v \mid \mathbf{Z}_t^v, \mathbf{Z}_0^v\right) = \mathcal{N}(\mathbf{Z}_{t-1}^v; \boldsymbol{\mu}_t\left(\mathbf{Z}_t^v, \mathbf{Z}_0^v\right), \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\mathbf{I}), \tag{4}$$

$$\boldsymbol{\mu}_t\left(\mathbf{Z}_t^v, \mathbf{Z}_0^v\right) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{Z}_t^v - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t), \tag{5}$$

where $\boldsymbol{\epsilon}_t$ represents the noise in $\mathbf{Z}_t^v$. Let $\boldsymbol{\epsilon}_\theta^v$ denotes the denoising network employed to estimate the noise $\boldsymbol{\epsilon}_t$ for view v, which is a U-Net architecture composed of multiple convolutional layers and ReLU activation functions. To encode the step information, we use sinusoidal and cosine positional encoding (Vaswani et al. 2017) to encode the time step t, namely,

$$\mathrm{PE}(t) = \left[\sin\left(\frac{t}{10000^{\frac{2i}{d}}}\right), \cos\left(\frac{t}{10000^{\frac{2i+1}{d}}}\right)\right], \tag{6}$$

where d is the encoding dimension, and $i = 0, 1, \ldots, \frac{d}{2} - 1$. The encoded time step vector is directly fed into the denoising network $\boldsymbol{\epsilon}_\theta^v$ to help the model understand the different stages of the generation process.

The objective of the diffusion model is to optimize the parameters θ of the denoising network to make the estimated noise vector $\epsilon_\theta^v(\mathbf{Z}_t^v, \mathrm{PE}(t))$ close to fully Gaussian noise $\boldsymbol{\epsilon}$ (Ho, Jain, and Abbeel 2020), namely,

$$\mathcal{L}_{diff} = \sum_{v=1}^{V} \mathbb{E}_{\mathbf{Z}_0^v, t, \epsilon}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta^v(\mathbf{Z}_t^v, \mathrm{PE}(t))\|^2\right]. \tag{7}$$

After the reverse denoising process, we achieved intra-view data generation. To achieve cross-view data generation for missing views through our diffusion model, we employ contrastive learning (Chen et al. 2020) on a small number of paired samples to align the generated views with the real views. Specifically, we conduct the contrastive loss between the generated view $\hat{\mathbf{Z}}^m$ and all other existing views:

$$\mathcal{L}_{gcl} = \frac{1}{2}\sum_{m=1}^{V}\sum_{n \neq m} \ell_{cl}^{(mn)}, \tag{8}$$

where the contrastive learning between generated view $\hat{\mathbf{Z}}^m$ and existing view $\mathbf{Z}^n$ is given as:

$$\ell_{cl}^{(mn)} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{sim(\hat{\mathbf{Z}}_i^m, \mathbf{Z}_i^n)/\tau_F}}{\sum_{j=1}^{N}\sum_{v=m,n} e^{sim(\hat{\mathbf{Z}}_i^m, \mathbf{Z}_j^v)/\tau_F}}, \tag{9}$$

where $sim(\cdot, \cdot)$ denotes the cosine similarity, τ_F is the temperature parameter that controls the softness.

To sum up, in the training stage, we employ $\mathcal{L}_{\mathbf{D}}$ to jointly train the entire module:

$$\mathcal{L}_{\mathbf{D}} = \mathcal{L}_{diff} + \mathcal{L}_{gcl}. \tag{10}$$

Data Recovery Process. During the inference stage, we generate the missing i-th view representations from the other existing views, using the denoising network ϵ_θ^i and iteratively applying the reverse denoising step $p_\theta(\mathbf{Z}_{t-1}^v \mid \mathbf{Z}_t^v)$ from $t = T_{ext}$ to $t = 1$ on each view. Formally,

$$\bar{\mathbf{Z}}_{t-1}^j = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{Z}_t^j - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta^i(\mathbf{Z}_t^j, \mathrm{PE}(t))\right) + \sigma_t\xi, \tag{11}$$

$$\hat{\mathbf{Z}}^i = \frac{1}{V - 1}\sum_{j \neq i}^{V} \bar{\mathbf{Z}}_0^j \tag{12}$$

where $\mathbf{Z}^j$ denotes the representations of the other existing views, $\hat{\mathbf{Z}}^i$ denotes the generated missing representation of the i-th view, σ_t denotes the noise standard deviation associated with the time step t, $\xi \sim \mathcal{N}(0, \mathbf{I})$ denotes a noise term, and $t = (T_{ext}, \ldots, 1)$.

Remark. Benefiting from the periodicity of the sinusoidal and cosine positional encoding (Vaswani et al. 2017), we can extrapolate the diffusion step during the inference stage (*i.e.*, $T_{ext} > T$), leading to further clustering benefits.

Instance-level Interactive Learning

To fully exploit the cross-view consistency and complementarity of samples, we designed an instance-level interactive learning module that consists of two parts: an attention fusion module and a view alignment module. Directly concatenating or averaging these features usually leads to degraded clustering performance. To fully leverage complementary information from different views, we adopted an attention fusion module that automatically perceives fusion weights of views, which guide the modules in the network to reinforce each other (Zhou and Shen 2020). Specifically, the attention fusion module consists of three fully connected layers followed by a softmax layer. It takes the concatenated feature $\mathbf{Z}$ as input and produces a view weight vector $\mathbf{w}$ as output, namely,

$$\mathbf{Z} = \left[\mathbf{Z}^1, \mathbf{Z}^2, \ldots, \mathbf{Z}^V\right] \tag{13}$$

$$\mathbf{w} = \mathrm{Softmax}\left(\mathrm{sigmoid}(\mathrm{MLP}(\mathbf{Z}))/\delta\right), \tag{14}$$

where $\mathrm{MLP}(\cdot)$ denotes three fully connected layers, and δ is a calibration factor. The Sigmoid function and calibration factor are used to prevent assigning scores close to one to the most relevant view features. Subsequently, the fused common representation $\mathbf{H} \in \mathbb{R}^{N \times d_v}$ can be obtained by $\mathbf{H} = \sum_{v=1}^{V} \mathbf{w}_v \mathbf{Z}^v$.

Through the attention fusion module, the complementary information in multi-view data is fully utilized, which in turn guides the view-specific representation learning. Therefore, we maximize consistency between the common representation $\mathbf{H}$ and the view-specific representation $\mathbf{Z}^v$ by employing the following maximum mutual information loss (Lin et al. 2022), namely,

$$\mathcal{L}_{\mathbf{I}} = -\sum_{v=1}^{V} I\left(\mathbf{H}; \mathbf{Z}^v\right) = -\sum_{v=1}^{V}\sum_{i,j} p\left(\mathbf{h}_i, \mathbf{z}_j^v\right)\log\frac{p\left(\mathbf{h}_i, \mathbf{z}_j^v\right)}{p\left(\mathbf{h}_i\right)p\left(\mathbf{z}_j^v\right)}. \tag{15}$$

where $I(\cdot)$ denotes the mutual information. By minimizing Eq. (15), the information correlation between the common representation and the view-specific representation among multiple views can be enhanced to effectively mine the cross-view instance consistency.

Category-level Interactive Learning

To obtain soft cluster assignments for end-to-end clustering, we design a shared classifier $g(\cdot)$ with a softmax layer for all views to obtain the soft cluster assignments, i.e., $\mathbf{Q}^v = g(\mathbf{Z}^v)$, where $\mathbf{Q}^v \in \mathbb{R}^{N \times K}$. Due to the heterogeneous view-private information, and to improve the model's robustness, we adopt category-level contrastive loss (Chen et al. 2020) to achieve the cluster consistency. We define the category-level contrastive loss between $\mathbf{Q}^m$ and $\mathbf{Q}^n$ as follows:

$$\ell_{sc}^{(mn)} = -\frac{1}{K}\sum_{j=1}^{K}\log\frac{e^{sim(\mathbf{Q}_{.j}^m, \mathbf{Q}_{.j}^n)/\tau_C}}{\sum_{k=1}^{K}\sum_{v=m,n} e^{sim(\mathbf{Q}_{.j}^m, \mathbf{Q}_{.k}^v)/\tau_C}}, \tag{16}$$

where τ_C is the temperature parameter that controls the softness. For all views, the category-level contrastive loss can be denoted as:

$$\mathcal{L}_{ccl} = \frac{1}{2}\sum_{m=1}^{V}\sum_{n\neq m}\ell_{sc}^{(mn)} + \sum_{v=1}^{V}\sum_{j=1}^{K} s_j^v \log s_j^v, \tag{17}$$

where $s_j^v = \frac{1}{N}\sum_{i=1}^{N} q_{ij}^v$ is the regularization term, which is used to avoid all samples being assigned to the same cluster (Huang, Gong, and Zhu 2020).

To further enhance the confidence of soft cluster assignments, we adopted KL divergence to guide complementary information learning through self-supervision, so that high-confidence instances not only improve their own representation learning, but also enhance the soft cluster assignments of other instances. Specifically, we similarly input the common representation $\mathbf{H}$ into the shared classifier to obtain the corresponding soft cluster assignments, i.e., $\mathbf{Q} = g(\mathbf{H})$. By taking the maximum soft cluster assignment for each instance in $\mathbf{Q}$ and $\mathbf{Q}^v$, we can obtain high-confidence soft cluster assignment:

$$\mathbf{q}_{ij} = \max\{\mathbf{Q}_{ij}, \mathbf{Q}_{ij}^v\}, \tag{18}$$

The high-confidence soft cluster assignment is used as the target assignment:

$$\mathbf{p}_{ij} = \frac{\mathbf{q}_{ij}^2}{\sum_{j=1}^{k}\mathbf{q}_{ij}^2}, \tag{19}$$

Then, the clustering results are optimized by applying KL divergence to enhance the soft cluster assignment with high confidence and further blur the instances near the cluster boundaries:

$$\mathcal{L}_{kl} = KL(\mathbf{P}\|\mathbf{Q}) = \sum_{i=1}^{N}\sum_{j=1}^{K}\mathbf{p}_{ij}\log\frac{\mathbf{p}_{ij}}{\mathbf{q}_{ij}}, \tag{20}$$

The overall loss of the category-level interactive learning module is:

$$\mathcal{L}_{\mathbf{C}} = \mathcal{L}_{ccl} + \mathcal{L}_{kl}. \tag{21}$$

The Objective Function

Our model is an end-to-end clustering method that does not require k-means (Bauckhage 2015) clustering to obtain the final clustering results. Therefore, we can optimize the entire model simultaneously. The total loss is as follows:

$$\mathcal{L} = \mathcal{L}_{\mathbf{R}} + \lambda_1\mathcal{L}_{\mathbf{D}} + \lambda_2\mathcal{L}_{\mathbf{I}} + \lambda_3\mathcal{L}_{\mathbf{C}}. \tag{22}$$

In our experiments, these three trade-off coefficients λ_1, λ_2, and λ_3 are all set to 1.

Experiments

To evaluate the effectiveness of DCG, we conducted extensive experiments to answer the following questions: (Q1) Does DCG outperform state-of-the-art IMVC methods? (Q2) Does each component of DCG contribute to the overall performance? (Q3) Do the hyperparameters affect the performance of DCG? (Q4) What is the clustering structure revealed by DCG?

Experimental Setup

We conducted experiments on five multi-view datasets, including Synthetic3D (Kumar, Rai, and Daume 2011), CUB (Wah et al. 2011), HandWritten (Asuncion and Newman 2007), LandUse-21 (Yang and Newsam 2010), and Fashion (Xiao, Rasul, and Vollgraf 2017). To evaluate the performance of handling incomplete multi-view data, following (Lin et al. 2021), we randomly select m instances and randomly delete one view, where the missing rate is m/n, with n representing the total number of instances. For a comprehensive analysis, three widely-used clustering metrics, including Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI), are used. A higher value of these metrics indicates better clustering performance. Detailed descriptions of the datasets and implementation are in the supplementary materials.

Comparisons with State of the Arts (Q1)

We compared DCG with eight state-of-the-art IMVC methods, including DCP (Lin et al. 2022), DSIMVC (Tang and Liu 2022), GCFAgg (Yan et al. 2023), CPSPAN (Jin et al. 2023), APADC (Xu et al. 2023), ProImp (Li et al. 2023), DVIMVC (Xu et al. 2024a), and ICMVC (Chao, Jiang, and Chu 2024). A detailed description of these baselines is provided in the supplementary materials.

We evaluated DCG and baselines with different missing rates. Table 1 reports the average clustering performance under five random experiments. It can be observed that: (1) Our DCG method outperforms all baselines methods in all cases. In particular, when the missing rate is 0.1 on CUB, DCG achieves 5.56%, 6.32%, and 10.56% improvements on ACC, NMI, and ARI, respectively, compared to the second best method. These results demonstrate the effectiveness and superiority of our proposed method over the baselines. (2) As the missing rates increase, the performance of all methods generally declines, indicating the missing data hinders the sufficient exploration of multi-view data. However, compared with other approaches, our DCG demonstrates greater robustness against increasing missing rates in most cases. For example, on HandWritten, when the missing rate increase from 0.1 to 0.7, the accuracy drop of DCG is 3%, while that of ICMVC is 8.88%, which demonstrates the stability of DCG.

Ablation Studies (Q2)

To evaluate the effectiveness of each component in our proposed DCG method, we conducted ablation experiments on CUB, HandWritten, and Fashion. Specifically, we designed several variants: "(*w/o*) $diff$", "(*w/o*) *gcl*", and "(*w/o*) $diff$&*gcl*," which respectively represent the removal of the diffusion model, the removal of the contrastive learning loss, and the removal of both. Moreover, "(*w/o*) **I**" indicates the removal of the maximum mutual information loss. Similarly, "(*w/o*) *ccl*", "(*w/o*) *kl*", and "(*w/o*) *ccl*&*kl*" respectively represent the removal of the category-level contrastive loss, the removal of the KL divergence loss, and the removal of both.

As shown in Table 2, we can draw the following conclusions: 1) The model performance decreases when any mod-

Dataset	Missing rates	0.1			0.3			0.5			0.7		
	Evaluation metrics	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
Synthetic3d	DCP (Lin et al. 2022)	88.00	65.17	67.87	79.83	55.21	56.69	85.53	58.33	62.56	81.50	52.75	53.39
	DSIMVC (Tang and Liu 2022)	73.11	59.02	58.38	70.78	56.41	55.43	67.44	51.74	49.42	66.33	48.96	46.36
	GCFAgg (Yan et al. 2023)	72.52	57.87	56.68	70.23	55.51	53.34	69.12	53.21	52.03	67.95	50.50	49.12
	CPSPAN(Jin et al. 2023)	88.83	65.51	69.12	87.50	61.79	66.68	86.33	59.29	63.65	82.33	52.24	54.59
	APADC (Xu et al. 2023)	85.73	59.07	62.98	86.47	60.84	64.90	84.38	58.48	61.85	80.83	49.70	53.42
	ProImp (Li et al. 2023)	86.55	61.79	65.42	85.50	59.29	62.81	82.78	54.76	57.69	76.67	44.88	46.42
	DVIMVC (Xu et al. 2024a)	50.03	28.43	24.57	46.97	22.07	15.93	50.32	25.64	18.79	58.28	36.84	30.77
	ICMVC (Chao, Jiang, and Chu 2024)	85.03	58.30	61.96	87.17	61.20	66.29	84.20	55.33	59.60	70.77	42.02	40.71
	DCG (Ours)	**91.23**	**71.35**	**76.05**	**88.00**	**64.11**	**68.48**	**87.67**	**63.23**	**67.52**	**85.50**	**57.13**	**62.17**
CUB	DCP (Lin et al. 2022)	42.77	55.42	33.39	40.60	52.37	31.41	38.87	50.15	31.18	38.20	47.97	29.18
	DSIMVC (Tang and Liu 2022)	63.67	59.85	46.23	49.22	49.93	32.80	49.89	49.61	32.35	36.78	40.30	23.72
	GCFAgg (Yan et al. 2023)	67.67	64.14	51.14	62.72	60.27	45.03	59.63	55.95	39.91	39.15	41.35	22.95
	CPSPAN(Jin et al. 2023)	76.67	71.38	58.65	74.33	70.33	56.84	73.33	69.68	57.17	68.00	68.36	53.28
	APADC (Xu et al. 2023)	53.04	59.05	42.98	51.43	59.06	40.52	48.53	59.93	41.54	42.57	50.49	31.91
	ProImp (Li et al. 2023)	69.33	69.65	55.93	71.83	64.45	51.64	69.56	64.19	52.06	61.28	56.37	42.34
	DVIMVC (Xu et al. 2024a)	63.83	63.00	50.50	70.42	68.43	56.35	68.98	64.81	52.61	55.53	52.55	39.54
	ICMVC (Chao, Jiang, and Chu 2024)	32.33	33.36	16.85	45.13	41.32	25.11	41.70	37.78	21.62	39.30	37.08	21.06
	DCG (Ours))	**82.23**	**77.70**	**69.21**	**77.17**	**71.35**	**59.85**	**75.50**	**72.21**	**59.12**	**74.67**	**70.19**	**56.41**
HandWritten	DCP (Lin et al. 2022)	63.66	70.44	52.82	75.68	79.05	67.45	72.07	76.17	63.81	66.13	69.04	55.91
	DSIMVC (Tang and Liu 2022)	71.37	70.24	60.30	67.47	65.27	54.26	56.90	57.99	44.12	49.27	48.02	34.10
	GCFAgg (Yan et al. 2023)	75.07	68.19	59.79	69.16	63.25	53.67	61.02	53.57	42.15	49.85	45.32	32.23
	CPSPAN(Jin et al. 2023)	68.70	69.06	59.17	82.20	75.05	69.58	72.55	68.60	57.51	76.15	69.04	61.33
	APADC(Xu et al. 2023)	72.58	71.29	59.07	60.66	64.84	48.75	52.70	63.12	39.96	53.68	55.29	36.20
	ProImp(Li et al. 2023)	81.05	78.48	70.40	80.60	77.26	69.72	78.98	74.08	66.47	76.20	69.63	61.10
	DVIMVC(Xu et al. 2024a)	29.08	19.36	11.06	26.63	16.57	8.84	25.13	15.69	8.09	24.52	14.25	7.71
	ICMVC(Chao, Jiang, and Chu 2024)	82.70	81.06	74.67	82.27	79.90	72.81	75.04	71.89	63.29	73.82	70.15	60.76
	DCG (Ours)	**82.75**	**82.63**	**74.88**	**82.70**	**80.54**	**73.96**	**80.80**	**76.21**	**70.45**	**79.75**	**74.00**	**67.54**
LandUse-21	DCP (Lin et al. 2022)	26.19	31.20	13.02	25.48	30.18	11.08	21.52	26.32	11.11	22.18	27.00	10.13
	DSIMVC (Tang and Liu 2022)	16.56	16.40	4.32	16.46	16.57	4.35	16.70	16.84	4.50	16.22	15.76	4.12
	GCFAgg (Yan et al. 2023)	19.05	19.99	6.32	18.53	19.89	6.24	18.43	19.47	5.98	19.05	20.33	6.37
	CPSPAN(Jin et al. 2023)	20.05	27.20	8.14	18.05	25.75	6.90	20.38	26.99	8.05	23.52	26.33	10.31
	APADC (Xu et al. 2023)	20.92	26.74	7.97	19.99	24.40	7.44	19.15	23.91	7.08	18.97	21.44	6.78
	ProImp (Li et al. 2023)	24.43	29.22	11.44	24.63	28.45	11.78	24.45	27.43	10.94	23.86	27.05	10.82
	DVIMVC (Xu et al. 2024a)	13.90	22.48	2.66	13.24	23.21	2.74	12.86	19.67	2.54	12.57	19.66	2.57
	ICMVC (Chao, Jiang, and Chu 2024)	26.81	30.71	13.60	26.12	29.83	12.87	24.43	27.27	11.12	22.97	25.15	9.65
	DCG (Ours)	**27.52**	**31.36**	**14.57**	**27.33**	**32.09**	**14.47**	**25.76**	**29.07**	**13.17**	**25.14**	**27.23**	**11.85**
Fashion	DCP (Lin et al. 2022)	83.70	84.30	76.50	71.80	70.90	52.50	60.80	59.50	33.10	49.90	48.40	19.10
	DSIMVC (Tang and Liu 2022)	88.00	86.40	81.10	87.30	85.00	78.90	83.50	80.30	73.70	75.71	71.53	69.00
	GCFAgg (Yan et al. 2023)	78.21	74.50	66.28	76.34	72.53	63.98	74.47	69.83	60.37	72.47	67.93	57.98
	CPSPAN(Jin et al. 2023)	66.16	68.45	55.73	64.80	68.22	55.55	54.81	64.13	48.84	66.32	68.44	55.99
	APADC (Xu et al. 2023)	81.40	86.50	73.30	80.90	85.01	73.10	75.40	81.50	67.60	52.90	59.78	37.40
	ProImp (Li et al. 2023)	92.88	88.34	86.09	74.11	76.97	66.42	89.76	81.94	79.93	76.28	74.44	66.71
	DVIMVC (Xu et al. 2024a)	79.38	80.22	71.50	82.26	79.82	72.46	80.17	77.09	69.51	76.95	74.25	65.52
	ICMVC (Chao, Jiang, and Chu 2024)	92.41	87.05	85.11	89.31	83.06	79.93	79.37	74.44	68.46	72.17	68.45	59.88
	DCG (Ours)	**95.83**	**91.29**	**91.19**	**93.13**	**86.99**	**86.00**	**90.04**	**82.25**	**79.99**	**85.76**	**76.42**	**72.79**

Table 1: Clustering results of all methods on five datasets with different missing rates. The best and second-best results are highlighted in bold and underlined, respectively.

ule in the DCG is removed, indicating that each module contributes significantly to improving the overall performance; 2) Removing both the diffusion model and the contrastive learning loss simultaneously results in the largest decrease in model performance, suggesting that our diffusion contrastive generation module significantly contributes to recovering missing view representations.

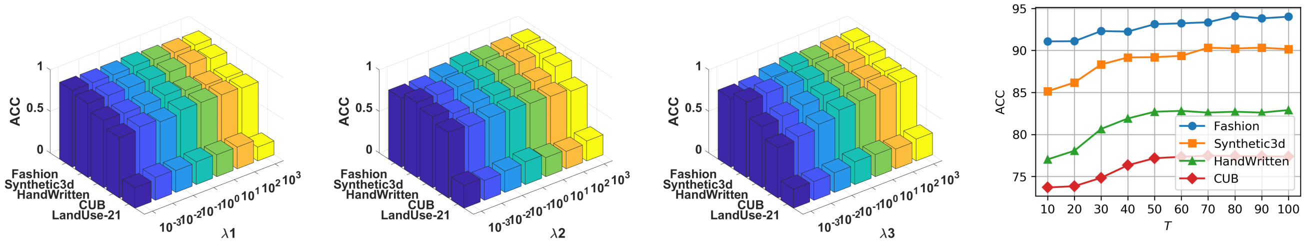
To further evaluate the effectiveness of the diffusion contrastive generation module, we replaced it with GAN (Wang et al. 2021) and Prediction (Lin et al. 2022), as shown in Table 3. The results show that our recovery strategy outperforms other methods, primarily because it innovatively combines diffusion processes with contrastive learning, enabling effective view generation and discriminative learning. Additionally, to verify the effectiveness of our data usage strategy, we changed the strategy to use only paired data for experiments. As shown in Table 3, the results show that, compared to the strategy of using only paired data, our data usage strategy can more effectively utilize the multi-view data, thereby improving clustering performance.

Hyperparameters Analysis (Q3)

To evaluate the robustness of our proposed method DCG to hyperparameters, we conducted experiments on different

Variants	Components					CUB			HandWritten			Fashion		
	$\mathcal{L}_{diff}$	$\mathcal{L}_{gcl}$	$\mathcal{L}_{\mathbf{I}}$	$\mathcal{L}_{ccl}$	$\mathcal{L}_{kl}$	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
(*w/o*) $diff$	×	✓	✓	✓	✓	70.96	67.35	53.49	74.65	74.13	63.44	81.69	80.68	72.57
(*w/o*) gcl	✓	×	✓	✓	✓	71.46	67.38	53.87	75.95	75.36	65.99	83.41	81.10	75.63
(*w/o*) $diff\&gcl$	×	×	✓	✓	✓	60.73	61.21	42.23	66.00	66.41	54.82	75.62	73.10	64.65
(*w/o*) $\mathbf{I}$	✓	✓	×	✓	✓	72.83	68.25	56.78	77.11	74.65	66.78	85.59	84.21	77.56
(*w/o*) ccl	✓	✓	✓	×	✓	71.23	66.78	53.78	73.72	72.65	62.78	80.56	78.33	72.89
(*w/o*) kl	✓	✓	✓	✓	×	73.63	67.98	56.53	79.23	77.82	71.63	87.80	83.22	79.46
(*w/o*) $ccl\&kl$	✓	✓	✓	×	×	68.89	65.34	51.29	70.79	70.23	61.66	78.57	76.58	70.28
Full model	✓	✓	✓	✓	✓	**77.17**	**71.35**	**59.85**	**82.70**	**80.54**	**73.96**	**93.13**	**86.99**	**86.00**

Table 2: Ablation study on CUB, HandWritten, and Fashion with the missing rate of 0.3.



(a) ACC vs. $\lambda 1$

(b) ACC vs. $\lambda 2$

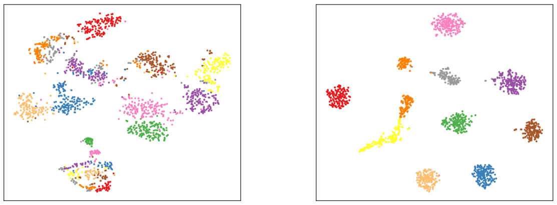
(c) ACC vs. $\lambda 3$

(d) ACC vs. T

Figure 3: Parameter sensitivity analysis on different datasets with the missing rate of 0.3.

Strategies	ACC	NMI	ARI
GAN (Wang et al. 2021)	55.82	53.43	48.39
Prediction (Lin et al. 2022)	71.13	66.76	54.61
Paired	72.83	66.80	53.98
Default	**77.17**	**71.35**	**59.85**

Table 3: Ablation study on different data recovery and data usage strategies on CUB with the missing rate of 0.3. "Paired" denotes use only paired data.



(a) Raw Features

(b) Final Result

Figure 4: t-SNE visualization of HandWritten with the missing rate of 0.3.

datasets with missing rate of 0.3 to analyze the impact of the three trade-off coefficients λ_1, λ_2, and λ_3, and the total number of time steps T on the clustering performance of DCG. Figure 3 shows the ACC of our method DCG as the trade-off coefficients vary from 10^{-3} to 10^{3} and the total number of time steps T varies from 10 to 100. The results indicate that DCG's clustering performance is insensitive to changes in λ_1, λ_2, and λ_3 within the range of 10^{-1} to 10^{1}. Additionally, the ACC gradually increases by extrapolating the diffusion step from 50 to 100, which validates our findings on the consistency between diffusion processes and data clustering.

Visualization Analysis (Q4)

To intuitively demonstrate the superiority of DCG, we use t-SNE to visualize both the raw features and the latent representations learned by DCG, as shown in Figure 4. We observe that after training our model, the instances from the same cluster become more compact, and the instances of different clusters are separated further apart. This indicates that our DCG method can effectively leverage the consistency and complementarity of information in multi-view data.

Conclusion

In this paper, we find that the diffusion process not only helps recover missing views but also enhances the compactness of data clustering as the diffusion progresses. Based on this phenomenon, we propose a novel Diffusion Contrastive Generation (DCG) method, which innovatively combines the diffusion process with contrastive learning, thereby enabling effective view generation and discriminative learning using limited paired data. Moreover, to further improve clustering performance, DCG integrates instance-level and category-level interactive learning to fully exploit the consistent and complementary information in multi-view data, achieving end-to-end clustering. Extensive experimental results demonstrate the effectiveness and superiority of our method on the IMVC task.

Acknowledgments

This research work is supported by the Big Data Computing Center of Southeast University.

References

Asuncion, A.; and Newman, D. 2007. UCI machine learning repository.

Bauckhage, C. 2015. K-means clustering is matrix factorization. *arXiv preprint arXiv:1512.07548*.

Cai, J.; Zhang, Y.; Fan, J.; Du, Y.; and Guo, W. 2024a. Dual contrastive graph-Level clustering with multiple cluster perspectives alignment. In *International Joint Conference on Artificial Intelligence*.

Cai, Y.; Zhang, Z.; Ghamisi, P.; Ding, Y.; Liu, X.; Cai, Z.; and Gloaguen, R. 2022. Superpixel contracted neighborhood contrastive subspace clustering network for hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–13.

Cai, Y.; Zhang, Z.; Liu, X.; Ding, Y.; Li, F.; and Tan, J. 2024b. Learning Unified Anchor Graph for Joint Clustering of Hyperspectral and LiDAR Data. *IEEE Transactions on Neural Networks and Learning Systems*.

Chao, G.; Jiang, Y.; and Chu, D. 2024. Incomplete Contrastive Multi-View Clustering with High-Confidence Guiding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 11221–11229.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, Y. 2022. Multi-omics clustering based on dual contrastive learning for cancer subtype identification. In *Proceedings of the 3rd International Symposium on Artificial Intelligence for Medicine Sciences*, 106–110.

Cui, C.; Ma, Y.; Cao, X.; Ye, W.; and Wang, Z. 2024. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 902–909.

Guan, R.; Li, Z.; Tu, W.; Wang, J.; Liu, Y.; Li, X.; Tang, C.; and Feng, R. 2024. Contrastive multi-view subspace clustering of hyperspectral images based on graph convolutional networks. *IEEE Transactions on Geoscience and Remote Sensing*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huang, J.; Gong, S.; and Zhu, X. 2020. Deep semantic clustering by partition confidence maximisation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8849–8858.

Jin, J.; Wang, S.; Dong, Z.; Liu, X.; and Zhu, E. 2023. Deep incomplete multi-view clustering with cross-view partial sample and prototype alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11600–11609.

Ke, G.; Wang, B.; Wang, X.; and He, S. 2024. Rethinking Multi-view Representation Learning via Distilled Disentangling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26774–26783.

Kumar, A.; Rai, P.; and Daume, H. 2011. Co-regularized multi-view spectral clustering. *Advances in neural information processing systems*, 24.

Li, G.; Rao, C.; Mo, J.; Zhang, Z.; Xing, W.; and Zhao, L. 2024a. Rethinking diffusion model for multi-contrast mri super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11365–11374.

Li, H.; Li, Y.; Yang, M.; Hu, P.; Peng, D.; and Peng, X. 2023. Incomplete multi-view clustering via prototype-based imputation. *arXiv preprint arXiv:2301.11045*.

Li, H.; Shen, C.; Torr, P.; Tresp, V.; and Gu, J. 2024b. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12006–12016.

Lin, Y.; Gou, Y.; Liu, X.; Bai, J.; Lv, J.; and Peng, X. 2022. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4447–4461.

Lin, Y.; Gou, Y.; Liu, Z.; Li, B.; Lv, J.; and Peng, X. 2021. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11174–11183.

Lin, Y.; Yang, M.; Yu, J.; Hu, P.; Zhang, C.; and Peng, X. 2023. Graph matching with bi-level noisy correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*, 23362–23371.

Lin, Y.; Zhang, J.; Huang, Z.; Liu, J.; Wen, Z.; and Peng, X. 2024. Multi-granularity correspondence learning from long-term noisy videos. In *International conference on learning representations*.

Liu, H.; Li, B.; Wu, H.; Liang, H.; Huang, Y.; Li, Y.; Ghanem, B.; and Zheng, Y. 2023. Combating mode collapse via offline manifold entropy estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8834–8842.

Lu, Y.; Li, H.; Li, Y.; Lin, Y.; and Peng, X. 2024. A survey on deep clustering: from the prior perspective. *Vicinagearth*, 1(1): 4.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11461–11471.

Luo, Y.; and Yang, Z. 2024. DynGAN: Solving Mode Collapse in GANs with Dynamic Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sun, Y.; Qin, Y.; Li, Y.; Peng, D.; Peng, X.; and Hu, P. 2024. Robust multi-view clustering with noisy correspondence. *IEEE Transactions on Knowledge and Data Engineering*.

Tang, H.; and Liu, Y. 2022. Deep safe incomplete multi-view clustering: Theorem and algorithm. In *International Conference on Machine Learning*, 21090–21110.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Neural Information Processing Systems*, 6000–6010.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wan, X.; Xiao, B.; Liu, X.; Liu, J.; Liang, W.; and Zhu, E. 2024. Fast continual multi-view clustering with incomplete views. *IEEE Transactions on Image Processing*.

Wang, Q.; Ding, Z.; Tao, Z.; Gao, Q.; and Fu, Y. 2021. Generative partial multi-view clustering with adaptive fusion and cycle consistency. *IEEE Transactions on Image Processing*, 30: 1771–1783.

Wen, J.; Deng, S.; Wong, W.; Guoqing, C.; Chao, H.; Fei, L.; and Xu, Y. 2024. Diffusion-based Missing-view Generation With the Application on Incomplete Multi-view Clustering. In *International Conference on Machine Learning*.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xu, G.; Wen, J.; Liu, C.; Hu, B.; Liu, Y.; Fei, L.; and Wang, W. 2024a. Deep Variational Incomplete Multi-View Clustering: Exploring Shared Clustering Structures. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16147–16155.

Xu, J.; Li, C.; Peng, L.; Ren, Y.; Shi, X.; Shen, H. T.; and Zhu, X. 2023. Adaptive feature projection with distribution alignment for deep incomplete multi-view clustering. *IEEE Transactions on Image Processing*, 32: 1354–1366.

Xu, Y.; Zhao, Y.; Xiao, Z.; and Hou, T. 2024b. Ufogen: You forward once large scale text-to-image generation via diffusion gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8196–8206.

Yan, W.; Zhang, Y.; Lv, C.; Tang, C.; Yue, G.; Liao, L.; and Lin, W. 2023. Gcfagg: Global and cross-view feature aggregation for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19863–19872.

Yan, W.; Zhang, Y.; Tang, C.; Zhou, W.; and Lin, W. 2024. Anchor-Sharing and Clusterwise Contrastive Network for Multiview Representation Learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Yang, Y.; and Newsam, S. 2010. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 270–279.

Zhang, C.; Wei, J.; Wang, B.; Li, Z.; Chen, C.; and Li, H. 2023. Robust spectral embedding completion based incomplete multi-view clustering. In *Proceedings of the 31st ACM International Conference on Multimedia*, 300–308.

Zhou, R.; and Shen, Y.-D. 2020. End-to-end adversarial-attention network for multi-modal clustering. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 14619–14628.