Diffusion-based Missing-view Generation With the Application on Incomplete Multi-view Clustering

Jie Wen [1] **Shijie Deng** [1] **Waikeung Wong*** [2 3] **Guoqing Chao** [4] **Chao Huang** [5] **Lunke Fei** [6] **Yong Xu** [1]

Abstract

As a branch of clustering, multi-view clustering has received much attention in recent years. In practical applications, a common phenomenon is that partial views of some samples may be missing in the collected multi-view data, which poses a severe challenge to design the multi-view learning model and explore complementary and consistent information. Currently, most of the incomplete multi-view clustering methods only focus on exploring the information of available views while few works study the missing view recovery for incomplete multi-view learning. To this end, we propose an innovative diffusion-based missing view generation (DMVG) network. Moreover, for the scenarios with high missing rates, we further propose an incomplete multi-view data augmentation strategy to enhance the recovery quality for the missing views. Extensive experimental results show that the proposed DMVG can not only accurately predict missing views, but also further enhance the subsequent clustering performance in comparison with several state-of-the-art incomplete multi-view clustering methods.

1. Introduction

In recent years, multi-view data, representing the same object from different views, has played an important role in machine learning and data mining (Wang et al., 2022; Wen et al., 2023a; Zhou et al., 2023; Ren et al., 2021). For example, for face recognition, researchers designed more robust recognition systems based on multi-modal face images collected by different sensors, such as the visual camera and infrared camera (Kortli et al., 2020). For social media analysis, it has been proved that combining text, images, and user behavior data enables the model to understand user interactions and social trends better (Chandrasekaran et al., 2021). Audiovisual and time information are considered for acoustic event classification (Liu et al., 2023b). Similarly, multi-view data often yields better results in clustering tasks (Wang et al., 2023; Kang et al., 2021; Zhu et al., 2020; Zhou et al., 2020; Liu et al., 2023a; Ren et al., 2020).

Clustering is a classic problem in unsupervised learning (Liu et al., 2023c). Past research works on multi-view clustering generally focus on exploring the complementary and consistent information of multi-view data based on the assumption that all views of data points are fully collected (Du et al., 2023; Ren & Sun, 2020; Yang et al., 2023; Zhou & Du, 2023; Liang et al., 2020). These methods commonly ignore the scenarios with missing views and thus cannot be applied to incomplete multi-view clustering tasks. Considering the universality of missing views in practical applications (Wen et al., 2023b), we study incomplete multi-view clustering (IMVC) in this paper. For IMVC, researchers have also designed various methods (Zhang et al., 2021; Liu et al., 2021; Li et al., 2022b; Zhang et al., 2024; Wang et al., 2021). Generally speaking, with the increase of missing view rates or the decrease of paired view rates, the clustering performance of nearly all IMVC methods declines, some even drastically. Based on the idea that recovering missing views and then conducting clustering could potentially enhance the clustering performance and reduce the sensitivity to missing view rates, researchers proposed some generative IMVC methods, such as AIMC (Xu et al., 2021), GP-MVC (Wang et al., 2021), COMPLETER (Lin et al., 2021), and DCP (Lin et al., 2023). However, these methods integrate missing view recovery and clustering into a single framework, which may result in more uncertainty and even very poor performance on both tasks. A better approach is to decompose incomplete multi-view learning into two independent sub-tasks: missing view generation and multi-view learning (MVL).

[1]Shenzhen Key Laboratory of Visual Object Detection and Recognition, Harbin Institute of Technology, Shenzhen, China [2]School of Fashion and Textiles, The Hong Kong Polytechnic University, Hong Kong [3]Laboratory for Artificial Intelligence in Design, Hong Kong [4]School of Computer Science and Technology, Harbin Institute of Technology, Weihai, China [5]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen, China [6]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, China. Correspondence to: Waikeung Wong <calvin.wong@polyu.edu.hk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

This allows any missing-view generation method to be combined easily with any MVL method. In existing works, two representative studies on missing view generation are VIGAN (Shang et al., 2017) and CRA (Tran et al., 2017). VIGAN can only perform mutual generation for dual-view data and has a relatively complex model composed of multiple sub-networks, which makes model training challenging and prone to model collapse. CRA is composed of multiple repetitive residual autoencoders (RA). Although CRA shows flexibility in recovering missing views, its training process is complex and unstable.

To address the above issues, we proposed a new Diffusion-based Missing-view Generation (DMVG) network. As an unsupervised learning method, DMVG aims to learn the intrinsic connections between views and use these connections to generate missing views from available views. Then for the incomplete multi-view clustering tasks, we can subsequently perform existing MVC methods on the filled data. Specifically, DMVG treats the view to be generated as the target view and other views as conditional views. Then, it progressively adds noise to the target view. Once sufficiently noised, the target view is encoded as standard Gaussian noise. Furthermore, by leveraging the information from its conditional views, the model gradually restores the original data from the noised target view. In our paper, to strengthen the model's recovery quality under a high missing rate, a data augmentation method is designed for DMVG, termed DA-DMVG. Experiments on missing view generation and subsequent MVC show that DMVG not only excels in recovering missing views but also enhances the performance of subsequent MVC. The main contributions of this paper are summarized as follows:

(1) A novel diffusion-based missing view generation method, DMVG, is proposed. It is the first work to apply the diffusion model for the task of missing view generation. Unlike most models that only achieve mutual generation for dual views, DMVG fully exploits the information from all available views to guide the generation of missing views.

(2) A multi-view data augmentation method is proposed for the incomplete multi-view data with a high missing view rate. The proposed data augmentation method can enhance the ability of DMVG to learn the latent connections between conditional views and the target view, which improves the prediction effect of missing views, especially for cases with a high missing view rate.

2. Preliminaries

2.1. Diffusion Model

The diffusion model, Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020), is defined as a series of noise addition and removal processes. Given a sample $x(0) \sim q(x)$, noise is progressively added as follows:

$$x(t) = \sqrt{\alpha_t}x(t-1) + \sqrt{1-\alpha_t}\varepsilon, \tag{1}$$

where t represents the step of noise addition, $x(t)$ represents the result after adding t times noise. $\varepsilon \sim \mathcal{N}(0,1)$ is random noise from a standard Gaussian distribution. α_t controls the magnitude of noise added at step t.

The process of adding noise from x_0 to x_t can be effectively achieved by an equivalent process as follows:

$$x(t) = \sqrt{\bar{\alpha}_t}x(0) + \sqrt{1-\bar{\alpha}_t}\varepsilon, \tag{2}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$. The posterior distribution of the noised data $x(t)$ follows Equations (3) and (4):

$$\begin{aligned} &p\left(x(t)|x(t-1)\right) = \\ &\mathcal{N}\left(x(t); \sqrt{\alpha_t}x(t-1), (1-\alpha_t)I\right), \end{aligned} \tag{3}$$

$$p\left(x(t)|x(0)\right) = \mathcal{N}\left(x(t); \sqrt{\bar{\alpha}_t}x(0), (1-\bar{\alpha}_t)I\right). \tag{4}$$

If $t \to +\infty$, then $\bar{\alpha}_t \to 0$, *i.e.*, $p\left(x(t)|x(0)\right) \sim \mathcal{N}(0,1)$. Thus, the reverse process is gradually denoising from $x(T) \sim \mathcal{N}(0,1)$ back to $x(0)$. DDPM demonstrated that the denoising process also follows a Gaussian distribution formulated as follows:

$$p\left(x(t-1)|x(t)\right) = \mathcal{N}\left(x(t-1); \mu_t\left(x(t),t\right), \sigma_t^2 I\right), \tag{5}$$

where μ_t and σ_t are the mean and standard deviation of the Gaussian distribution, respectively, as shown in Equations (6) and (7):

$$\mu_t(x(t),t) = \frac{1}{\sqrt{\alpha_t}}\left(x(t) - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta\left(x(t),t\right)\right), \tag{6}$$

$$\sigma_t = \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t)}, \tag{7}$$

where ε_θ represents the noise predictor model to predict the noise ε added before. θ represents the model parameters. And the loss function of ε_θ is the mean square error between the predicted noise and the noise added before, as shown in Equation (8):

$$\mathcal{L} = \left\|\varepsilon - \varepsilon_\theta\left(x(t),t\right)\right\|_F^2. \tag{8}$$

Consequently, the denoising process is achieved by sampling from the distribution (5) as follows:

$$x(t-1) = \mu_t\left(x(t),t\right) + \sigma_t\varepsilon. \tag{9}$$

2.2. Conditional Diffusion Model

Diffusion models generate random samples, while the goal of missing-view generation is to predict missing-views

based on available views. This requires that the generated view match the other available views of the sample, rather than randomly generating a sample conforming to the data distribution of the target view. To achieve this, one should start with conditional diffusion models (Ho & Salimans, 2021; Dhariwal & Nichol, 2021). Unlike diffusion models where ε_θ is only related to $x(t)$ and t, ε_θ also depends on the condition c in conditional diffusion models, *i.e.*, $\varepsilon_\theta\left(x(t), t, c\right)$.

In conditional diffusion models, the condition c can take various forms, such as text, labels, images, and other features, which is integrated into each step of the model and influence the entire generating process. So, the denoising process of the conditional diffusion model follows a Gaussian distribution as follows:

$$\begin{aligned} &p\left(x(t-1)|x(t), c\right) = \\ &\mathcal{N}\left(x(t-1); \mu_t\left(x(t), t, c\right), \sigma_t^2 I\right), \end{aligned} \tag{10}$$

where the condition c is used to adjust the mean μ_t and standard deviation σ_t during the denoising process, making the generated data related not only to the denoising step but also to the condition c. Noting that since the standard deviation σ_t is only related to the predefined noise addition scheme α, condition c does not change the standard deviation during the denoising process. Therefore, the mean μ_t of the denoising process of the conditional diffusion model can be modified as follows:

$$\begin{aligned} &\mu_t\left(x(t), t, c\right) = \\ &\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta\left(x(t), t, c\right)\right). \end{aligned} \tag{11}$$

In the same way, the loss function of the conditional diffusion model can be modified as:

$$\mathcal{L}_{\text{cond}} = \left\|\varepsilon - \varepsilon_\theta\left(x(t), t, c\right)\right\|_F^2. \tag{12}$$

3. Diffusion-based Missing-view Generation

In the previous section, we have analyzed that the conditional diffusion model is more suitable for generating the target missing views under certain given conditions. And thus we attempt to address the missing view generation issue based on the conditional diffusion model. In this section, we will introduce our innovative Diffusion-based Missing-view Generation method, referred to as DMVG, and present how to apply conditional diffusion models to predict missing views in detail. Subsequently, we will further present a data augmentation method based on DMVG, termed DA-DMVG (Data-Augmented DMVG), for missing view generation with a high missing rate. The framework of DMVG and data augmentation is shown in Figure 1.

3.1. Motivation of the Missing-view Generation

Assume that $\mathbf{X} = \{X^{(1)}, X^{(2)}, ..., X^{(l)}\}$ represents incomplete multi-view data with l views, where $X^{(v)} = [x_1^{(v)}, x_2^{(v)}, ..., x_n^{(v)}] \in \mathbb{R}^{d_v \times n}$ denotes the data in the v-th view. n denotes the number of samples and d_v denotes the feature dimension of the v-th view. If the v-th view of sample i is missing, then $x_i^{(v)} = \emptyset$. Here, $\emptyset$ represents empty data, which is filled with a meaningless feature encoding during the training process, such as a zero vector. The task of missing-view generation is to predict/recover the missing information in the dataset. Since different views of a sample describe the same sample, there are some connections between these views. Once the model grasps the underlying associations of different views, it acquires the capability to generate missing views from the available ones.

3.2. Model Structure and Training Loss

Model Structure. our DMVG is based on a UNet architecture (Ronneberger et al., 2015), which incorporates cascading down-samplers and up-samplers, along with skip connections for efficient information flow. The condition view data is concatenated at the lowest layer of the down-sampling path, and the embedding of the time step t is integrated during both the down-sampling and up-sampling phases.

In our work, conditional diffusion models are apt for the task of generating missing views. The missing views will be treated as the generation target and other views are set as the conditions for the conditional diffusion model. For the dataset $\mathbf{X}$ containing l views, each sample potentially has some missing instances. Therefore, the designed diffusion model needs to be trained for each view, specifically for generating that view. For example, if we need to predict the v-th view, *i.e.*, the target view to generate is $x^{(v)}$, the corresponding conditional views $c^{(v)}$ will be $\left\{x^{(j)}\right\}_{j\neq v}$. We define that the model for generating view v is denoted as $\varepsilon_\theta^{(v)}$. Let $x^{(v)}(t)$ be the result of adding noise t times to the v-th view of the sample, as shown in Equation (13):

$$x^{(v)}(t) = \sqrt{\bar{\alpha}_t}x^{(v)}(0) + \sqrt{1-\bar{\alpha}_t}\varepsilon. \tag{13}$$

During the training process, sample pairs $\left(x^{(v)}, c^{(v)}\right)$ are required as training data to supervise the model to generate real data. Therefore, this method first selects all samples, whose v-th view is not missing, to serve as training data, enabling the model to grasp the intrinsic connection from conditional views $c^{(v)}$ to the target view $x^{(v)}$. After training, the model can predict samples whose v-th view is missing.

Since missing views are random and uncontrollable, the conditional view set $c^{(v)} = \left\{x^{(j)}\right\}_{j\neq v}$ may also have missing views. In our paper, we define three scenarios for condition $c^{(v)}$ according to the practical view-missing cases: Full Con-

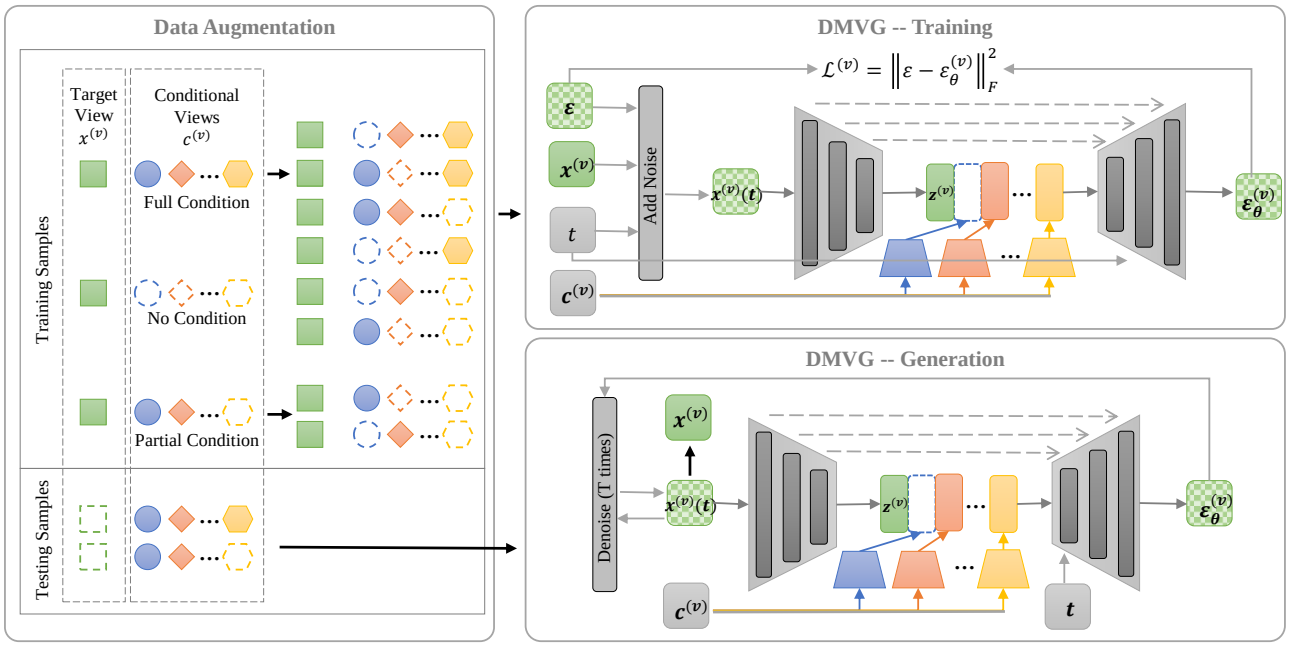


Figure 1. The framework of DMVG and data augmentation method. For data preprocessing, we initially select samples where the target view is not missing to serve as the training data. These training samples are then augmented as described in Section 3.4. During model training, a random noise is added to the target view, and the noisy target view along with all conditional views are fed into DMVG to fit this noise. In the model inference phase, the trained model is used to predict the missing samples of the target view. The detailed generation process is discussed in Section 3.3.

dition, No Condition, and Partial Condition. For the above three cases, the v-th view of samples should be available to train the generation model of the v-th view.

Full Condition. There are some complement samples with fully observed views. In this case, all available views except the target v-th view of those samples are used as the conditional view $c^{(v)}$ for model training.

No Condition. For some samples, only the target v-th view is available, and all of the other views are missing. For these samples, it is impossible to learn the mapping from conditional views $c^{(v)} = \left\{x^{(j)}\right\}_{j\neq v} = \emptyset$ to the target view $x^{(v)}$. However, Classifier-free Diffusion Guidance (CFG) (Ho & Salimans, 2021) pointed out that training diffusion models with unconditional samples can lead to lower loss and better generation effects, because unconditional training enables the model to learn the basic distribution of the data, preventing model bias by just learning the distribution under specific conditions. This helps the model understand the overall structure and features of the data better, improving its generalization performance and thus enhancing its performance in conditional generation tasks.

Partial Condition. Another case for incomplete data is that some incomplete samples have two observed views at least, including the v-th view. For missing-view generation model training, these samples are the most critical. Inspired by unconditional training, we believe that training with 'partial condition' is akin to training with 'no condition', helping the model learn the distribution under partial conditions. On one hand, this matches the task of predicting missing-views. On the other hand, it enhances the model's generalization ability like unconditional training.

To ensure the generalization ability, the model training should cover full, no, and partial conditions. Indeed, there is no need to deliberately differentiate these three types in the practical training process (simply set the missing parts of the conditional views $c^{(v)} = \left\{x^{(j)}\right\}_{j\neq v}$ to $\emptyset$). Furthermore, considering that unconditional training can be regarded as a regularization to prevent the model from over-relying on the condition, we also additionally train the model by setting the conditional views $c^{(v)}$ to $\emptyset$ with a certain probability p in addition to unconditional training caused by missing views. Based on the above analysis, we design our DMVG network with a unified structure as shown in the 'DMVG – Training' phase of Figure 1, which is an example for generating the v-th view of samples. For this sub-network corresponding to the v-th view, the inputs are $x^{(v)}(t)$, $c^{(v)} = \left\{x^{(j)}\right\}_{j\neq v}$, and t, and the output is the predicted noise $\varepsilon_\theta^{(v)}(x^{(v)}(t), t, c^{(v)})$. The backbone of the model is based on UNet, where the noised data $x^{(v)}(t)$ is encoded into a feature vector $z^{(v)}$ by the encoder of UNet, concatenated with the encoded condition $c^{(v)}$, and then decoded by the decoder with skip connections and noise step t.

Training Loss. The loss function $\mathcal{L}^{(v)}$ for the model $\varepsilon_\theta^{(v)}$

corresponding to the v-th view is designed as follows:

$$\mathcal{L}^{(v)} = \left\| \varepsilon - \varepsilon_\theta^{(v)} \left(x^{(v)}(t), t, c^{(v)} \right) \right\|_F^2. \tag{14}$$

3.3. Model Training and Missing View Generation

Referring to the loss function shown in Equation (14), we summarize the training process of DMVG in Algorithm 1. It is evident from the training process that the core idea of DMVG is to add noise to $x^{(v)}$ through Equation (13) and predict the noise added previously, based on the noisy data $x^{(v)}(t)$, the noise step t, and the conditional information $c^{(v)}$. The output of the model is merely a predicted noise. To predict the missing-view, it is necessary to progressively denoise $x^{(v)}(t)$ based on the predicted noise. The simplest denoising method is shown as follows:

$$x^{(v)}(t-1) = \mu_t \left(x^{(v)}(t), t, c^{(v)} \right) + \sigma_t \varepsilon, \tag{15}$$

where σ_t is the standard deviation as shown in Equation (7). μ_t is the mean calculated as follows:

$$\begin{gathered} \mu_t(x^{(v)}(t), t, c^{(v)}) = \\ \frac{1}{\sqrt{\alpha_t}} \left(x^{(v)}(t) - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \varepsilon_\theta^{(v)}(x^{(v)}(t), t, c^{(v)}) \right). \end{gathered} \tag{16}$$

To fully leverage the benefits of no condition training in enhancing the generation quality, following (Ho & Salimans, 2021), we further introduce a sampling method from the model $\varepsilon_\theta^{(v)}$ as follows:

$$\begin{gathered} \hat{\varepsilon}_\theta^{(v)} \left(x^{(v)}(t), t, c^{(v)} \right) = \\ (1+\omega)\varepsilon_\theta^{(v)} \left(x^{(v)}(t), t, c^{(v)} \right) - \omega \varepsilon_\theta^{(v)} \left(x^{(v)}(t), t \right), \end{gathered} \tag{17}$$

where $\varepsilon_\theta^{(v)}(x^{(v)}(t), t, c^{(v)})$ denotes conditional sampling, $\varepsilon_\theta^{(v)}(x^{(v)}(t), t)$ denotes unconditional sampling. $\omega \geq 0$ is a hyperparameter to control how much the generation process follows unconditional sampling. By adjusting ω, one can tune the degree of match between the generated results and the condition $c^{(v)}$ as well as the quality of generation. In our DMVG, $\omega = 1$.

Therefore, μ_t in the denoising process of DMVG is modified from Equation (16) to Equation (18):

$$\begin{gathered} \mu_t(x^{(v)}(t), t, c^{(v)}) = \\ \frac{1}{\sqrt{\alpha_t}} \left(x^{(v)}(t) - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \hat{\varepsilon}_\theta^{(v)}(x^{(v)}(t), t, c^{(v)}) \right). \end{gathered} \tag{18}$$

In summary, the missing-view generation process is illustrated in Algorithm 2. After generating the missing views by Algorithm 2, all conventional multi-view learning methods and the incomplete multi-view learning methods can be implemented on the new data composed of recovered missing views and original un-missing views.

Algorithm 1 Training of DMVG

input: incomplete multi-view data $\mathbf{X} = \left\{ X^{(v)} \right\}_{v=1}^{l}$, steps T, noise scheme $\{\alpha_t\}_{t=1}^{T}$, probability of no-conditional training p.
while not converged **do**
  Randomly sample a batch of x from X.
  Construct $x^{(v)}$ and $c^{(v)} = \left\{ x^{(j)} \right\}_{j \neq v}$ from x.
  Set $c^{(v)}$ to $\emptyset$ with probability p.
  Sample t from $[1, 2, ..., T]$.
  Sample ε from $\mathcal{N}(0, 1)$.
  Calculate $x^{(v)}(t)$ via Equation (13).
  Calculate $\mathcal{L}$ via Equation (14).
  Calculate gradients and update the model weights θ.
end while
output: noise predictor $\varepsilon_\theta^{(v)}$ for the v-th view.

Algorithm 2 Generation of DMVG

input: noise predictor $\varepsilon_\theta^{(v)}$ for the v-th view, conditional view $c^{(v)} = \left\{ x^{(j)} \right\}_{j \neq v}$, steps T, noise scheme $\{\alpha_t\}_{t=1}^{T}$, ω.
Sample $x^{(v)}(T)$ from $\mathcal{N}(0, 1)$.
for $t = T$ **to** 1 **do**
  Sample ε from $\mathcal{N}(0, 1)$ (if $t = 1$, $\varepsilon = 0$).
  Calculate $\varepsilon_\theta^{(v)} \left(x^{(v)}(t), c^{(v)} \right)$ and $\varepsilon_\theta^{(v)} \left(x^{(v)}(t) \right)$.
  Synthesize the noise $\hat{\varepsilon}_\theta^{(v)}$ via Equation (17).
  Calculate μ_t via Equation (18).
  Calculate σ_t via Equation (7).
  Calculate $x^{(v)}(t-1)$ via Equation (15).
end for
output: $x^{(v)} = x^{(v)}(0)$.

3.4. Data Augmentation for Incomplete Multi-view Data

For the proposed method, missing views are set to $\emptyset$ during the training process. However, this approach introduces a new challenge. When we use the model to generate the v-th view, if the missing rate is low, the model may identify one or a few views that are strongly correlated with the v-th view and focus more attention on these views. In other words, the performance of the generation model may highly depend on these few over-attended views with strong correlations with the target view. Therefore, for the incomplete multi-view data with a high missing rate, the model might struggle to produce good results when these few over-attended views strongly correlated with the target view are unavailable for generating their missing views.

To address the above challenge, we further propose an innovative data augmentation strategy shown in 'Data Augmentation' of Figure 1, and refer to the DMVG employing this strategy as Data Augmentation DMVG (DA-DMVG).

The core of this strategy lies in constructing pairs of samples with higher missing view rates from existing samples as supplemental training data. In detail, if there are m available views in a training sample, *i.e.*, $m-1$ available views in conditional views, we can create $2^{m-1}-2$ training samples additionally. Training the model with these augmented supplemental incomplete data enhances its ability to predict missing views, particularly in the cases with high missing view rate. This data augmentation method enables the model to learn more knowledge from limited conditional information, ensuring robust predictive performance across diverse data-missing scenarios. This strategy not only bolsters model robustness but also broadens its practical applicability, especially where conditional information is significantly lacking.

4. Experiments

4.1. Experiment Setup

Datasets. Caltech7 (Li et al., 2022a; 2015) consists of 1474 images of 7 kinds of objects, each with 6 feature views produced by Gabor, Wavelet moments, CENTRIST, Histogram of oriented gradients, GIST, and Local binary pattern feature extractors, respectively. **NH-face** (Cao et al., 2015) is derived from the film "Notting Hill" and contains 4660 facial images with 3 views represented by gray, gabor, and LBP features. **Multi-Modal CelebA-HQ** (Xia et al., 2021) contains 30000 pairs of human faces, including RGB and sketch images. Due to the high computational cost associated with applying diffusion to the original images, we trained an Autoencoder (AE) for each of the RGB and sketch modalities to compress the data to 512 dimensions. The network architecture of this AE is similar to UNet. Their key differences are the absence of skip connections, the embedding of step t, and the concatenating of the conditional view in the bottom of down-samplers. **Carl** (Espinosa-Duró et al., 2013) is a multi-view facial dataset with 2460 pairs, each including gray, infrared, and thermal images. As the same, we train 3 AEs to get their features of 512 dimensions, respectively.

Metrics. For missing-view generation experiments, following VIGAN and CRA, we select root mean squared error (RMSE), normalized mean squared error (NMSE), and peak signal-to-noise ratio (PSNR) as evaluation metrics. For clustering experiments, we follow popular experimental settings in (Wang et al., 2019; Wen et al., 2019) and select accuracy (ACC), normalized mutual information (NMI), and purity(PUR) as evaluation metrics.

Baselines. For missing-view generation experiments, we compared our method with two popular missing view generation methods (*i.e.*, VIGAN and CRA), and a method combined with missing view generation and clustering: DCP. For these baseline methods, we strived to adhere as closely as possible to the recommended parameter settings specified in the original literature. For clustering experiments, we compared four MVC methods and eight popular IMVC methods. The compared **MVC** methods are: Best Single View (BSV) (Zhao et al., 2016), Concat (Zhao et al., 2016), Multi-view Non-negative Matrix Factorization (MultiNMF) (Liu et al., 2013), and Co-regularized Multi-view Spectral Clustering (CCR-MVSC) (Kumar et al., 2011). For the compared **IMVC** methods, Doubly Aligned Incomplete Multi-view Clustering (DAIMC) (Hu & Chen, 2018) and One-Pass Incomplete Multi-View Clustering (OPIMC) (Hu & Chen, 2019) are based on matrix factorization. Efficient and Effective Incomplete Multi-View Clustering (EEIMVC) (Liu et al., 2019) and One-Stage Late Fusion Incomplete Multi-view Clustering (OSLF-IMVC) (Zhang et al., 2021) are kernel-based methods. Perturbation-oriented Incomplete multi-view Clustering (PIC) (Wang et al., 2019), Self-representation Subspace Clustering for Incomplete Multi-view (IMSR) (Liu et al., 2021), and Simultaneous Representation Learning and Clustering (SRLC) (Zhuge et al., 2019) are based on graph learning. Projective Incomplete Multi-view Clustering (PIMVC) (Deng et al., 2023) is a representative work based on projection learning.

Parameter Settings. In DMVG, we set $T = 1000$ and interpolate α_t in the range of $[1-10^{-6}, 1-2\times10^{-2}]$. The batch size is adapted based on the dataset size, with a learning rate as 10^{-4}. We utilize the Adam optimizer, with the learning rate linearly decreasing to zero over 1000 training epochs. As for the initialization, we employ a straightforward random initialization approach.The code of our DMVG is released at: `https://github.com/ckghostwj/DMVG/tree/main`.

4.2. Experiments on Missing-view Generation

For missing-view generation, we carried out experiments on Multi-Modal CelebA-HQ, Carl, and NH-face datasets.

4.2.1. Experimental Results

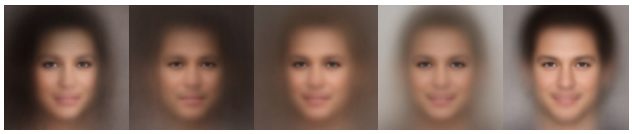
Experiment 1: Compare with VIGAN. We compare DMVG with VIGAN on Multi-Modal CelebA-HQ dataset with 0.1 missing instances in each view under the condition that each sample should have at least one view. Several missing view recovery results are shown in Figures 2 and 3, which intuitively demonstrate that DMVG successfully recovers the missing views while VIGAN fails because of mode collapse. More experimental results in terms of the RMSE, NMSE, and PSNR metrics are shown in Table 1, which demonstrates that DMVG significantly outperforms VIGAN in predicting RGB and sketch missing images on Multi-Modal CelebA-HQ dataset.

Experiment 2: Compare with CRA. We compare DMVG with CRA on Carl and NH-face datasets with the missing

(a) Original sketch images

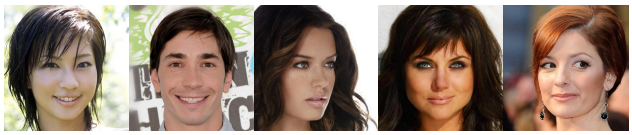
(b) Original RGB missing images (not used in model training)

(c) Generated RGB images via DMVG



(d) Generated RGB images via VIGAN

Figure 2. Generating RGB missing images according to their sketch images on Multi-Modal CelebA-HQ.



(a) Original RGB images

(b) Original sketch missing images (not used in model training)

(c) Generated sketch images via DMVG

(d) Generated sketch images via VIGAN

Figure 3. Generating sketch missing images according to their RGB images on Multi-Modal CelebA-HQ.

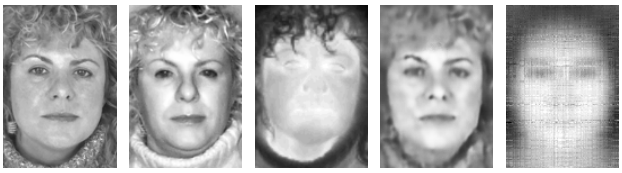
instance rate of 0.1, 0.3, and 0.5 for each view. The quantitative results in terms of RMSE, NMSE, and PSNR are shown in Tables 2 and 3, which also demonstrate that our method can get better recovery results than CRA for the missing views. To visually compare DMVG and CRA, we show 3 visualization examples of Carl with 0.1 missing views, where Figures 4, 5, and 6 are the recovered result of classic, infrared, and thermal image examples according to their corresponding available images from other modalities, respectively. We can observe that although the results of DMVG are a little blurry, almost all important information is recovered. Importantly, it is obvious that our recovery results look much better than CRA for the missing images.

Experiments 3: compare with DCP. DCP is a method combined with missing view generation and clustering. However, this limits its application to other MVL methods. On the other hand, comparing the experimental results of DCP and ours shown in Tables 1, 2 and 3, DCP gets poor generation quality. This may be attributed to its design only focusing on enhancing the clustering accuracy rather than precisely restoring missing views.

For more experiments on missing-view generation, please refer to Section A in the appendix.

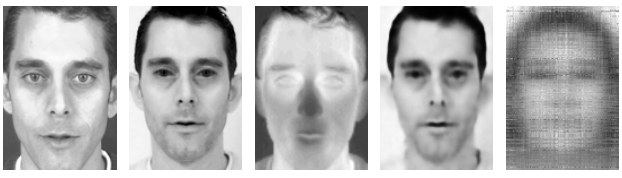
Table 1. RMSE, NMSE, and PSNR of DCP, VIGAN and DMVG on Multi-Modal CelebA-HQ with 10% missing views.

VIEW	METHOD	RMSE↓	NMSE↓	PSNR↑
3*RGB	DCP	0.83	2.05	1.64
	VIGAN	0.25	0.74	12.10
	DMVG	**0.24**	**0.70**	**12.34**
3*SKETCH	DCP	0.66	1.49	3.58
	VIGAN	0.33	1.50	9.57
	DMVG	**0.29**	**1.14**	**10.76**



(a) (b) (c) (d) (e)

Figure 4. Generating classic image according to the infrared and thermal images on Carl dataset. (a) Classic missing image, (b) infrared, (c) thermal, (d) generated classic image via DMVG, (e) generated classic image via CRA.



(a) (b) (c) (d) (e)

Figure 5. Generating infrared image according to classic and thermal images on Carl dataset. (a) Classic, (b) infrared missing image, (c) thermal, (d) generated infrared image via DMVG, (e) generated infrared image via CRA.

Table 2. RMSE, NMSE, and PSNR of DCP, CRA and DMVG on Carl with 10%, 30%, and 50% missing views.

VIEW	METHOD	RMSE↓			NMSE↓			PSNR↑		
		0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
3*CLASSIC	DCP	0.60	0.60	0.61	5.34	5.66	5.74	4.49	4.49	4.36
	CRA	0.16	0.17	0.18	0.36	0.47	0.53	16.18	15.30	14.68
	DMVG	**0.11**	**0.13**	**0.17**	**0.18**	**0.30**	**0.44**	**19.28**	**17.24**	**15.53**
3*INFRARED	DCP	0.57	0.60	0.65	5.92	6.91	8.09	4.82	4.45	3.78
	CRA	0.18	0.19	0.20	0.56	0.72	0.81	15.09	14.25	13.80
	DMVG	**0.12**	**0.14**	**0.15**	**0.26**	**0.33**	**0.45**	**18.37**	**17.70**	**16.35**
3*THERMAL	DCP	0.62	0.66	0.67	5.62	6.48	6.71	4.17	3.57	3.42
	CRA	0.14	0.15	0.17	0.27	0.32	0.41	17.35	16.60	15.53
	DMVG	**0.08**	**0.09**	**0.11**	**0.10**	**0.13**	**0.19**	**21.48**	**20.65**	**18.95**

Table 3. RMSE, NMSE, and PSNR of DCP, CRA, DMVG500, DMVG1000, and DA-DMVG1000 on NH-face with 10%, 30%, and 50% missing views.

VIEW	METHOD	RMSE↓			NMSE↓			PSNR↑		
		0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
5*GREY	DCP	0.32	0.29	0.29	2.75	2.51	2.62	9.92	10.88	10.70
	CRA	0.18	0.18	0.22	1.02	1.03	1.34	13.04	14.78	14.77
	DMVG500	0.06	0.08	0.17	0.11	0.18	0.88	24.02	22.43	15.46
	DMVG1000	0.06	0.08	0.12	0.10	0.20	0.44	24.51	21.87	18.45
	DA-DMVG500	**0.05**	**0.07**	**0.10**	**0.07**	**0.15**	**0.28**	**26.02**	**23.19**	**20.44**
5*GABOR	DCP	0.31	0.34	0.40	3.28	3.90	5.29	10.13	9.39	8.05
	CRA	0.19	0.19	0.20	1.21	1.24	1.36	13.93	14.49	14.37
	DMVG500	0.04	0.06	**0.08**	0.06	0.13	**0.19**	27.87	24.30	**22.40**
	DMVG1000	0.04	0.05	**0.08**	0.06	**0.07**	0.21	28.06	26.73	22.10
	DA-DMVG500	**0.03**	**0.04**	**0.08**	**0.04**	**0.07**	**0.19**	**29.30**	**27.01**	**22.40**
5*LBP	DCP	0.10	0.22	0.19	0.81	3.73	2.92	19.88	13.21	14.30
	CRA	0.11	0.28	0.38	1.01	6.39	11.33	18.89	10.91	8.39
	DMVG500	**0.07**	0.08	**0.08**	0.39	0.44	0.52	23.05	22.46	21.83
	DMVG1000	**0.07**	**0.07**	**0.08**	**0.38**	0.43	0.51	**23.15**	22.62	21.88
	DA-DMVG500	**0.07**	**0.07**	**0.08**	0.40	**0.41**	**0.48**	22.97	**22.79**	**22.16**

(a) (b) (c) (d) (e)

Figure 6. Generating thermal image according to classic and infrared images on Carl dataset. (a) Classic, (b) infrared, (c) thermal missing image, (d) generated thermal image via DMVG, (e) generated thermal image via CRA.

4.2.2. Ablation Study for Data Augmentation

To verify the effectiveness of the proposed data augmentation strategy, we compare three variants of our method, *i.e.*, DA-DMVG500, DMVG500, and DMVG1000 on the NH-face dataset with different missing rates. For DA-DMVG500, we applied data augmentation and trained the model for 500 epochs. For DMVG500 and DMVG1000, no data augmentation was used, and the model was trained for 500 epochs and 1000 epochs, respectively. Since data augmentation was employed in DA-DMVG500, the number of training samples increased, leading to more frequent updates of the model parameters under the same batch size. To compare the effects of data augmentation more fairly, DA-DMVG500 and DMVG1000 were analyzed side by side. Quantitative results in terms of RMSE, NMSE, and PSNR are shown in Table 3. We can observe that DA-DMVG is always better than DMVG, which confirms the positive impact of the proposed data augmentation strategy for missing view generation. In addition, Figure 7 shows several generated gray images from a single view, *i.e.*, LBP or GABOR, with 0.3 missing views. The results show that DA-DMVG can get clearer results, also indicating that the data augmentation strategy is effective in improving the recovery quality for the missing views.

4.3. IMVC Experiments after Missing-view Generation

To verify whether DA-DMVG is beneficial to the downstream tasks, we take the IMVC task as an example and compare the clustering performance with several state-of-

Table 4. ACC(%), NMI(%), and PUR(%) of different methods on Caltech7 with 10%, 30%, and 50% missing views.

	ACC			NMI			PUR		
METHOD	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
BSV	43.89	39.06	38.31	39.66	31.63	26.81	84.08	75.25	68.97
DAD+BSV	**50.05**	**50.65**	**49.06**	**46.53**	**45.59**	**42.13**	**87.67**	**84.53**	**85.60**
CONCAT	41.25	40.55	38.06	43.48	37.99	30.28	84.91	82.54	77.56
DAD+CONCAT	**44.76**	**44.60**	**43.72**	**46.43**	**44.77**	**43.58**	**86.05**	**85.18**	**84.97**
MULTINMF	46.39	40.61	37.92	30.16	28.52	28.88	79.80	78.43	78.03
DAD+MULTINMF	**50.80**	**51.64**	**45.69**	**35.77**	**34.10**	**32.48**	**81.28**	**80.30**	**78.09**
CCR-MVSC	38.47	38.16	37.35	35.53	34.18	31.52	78.33	77.89	76.21
DAD+CCR-MVSC	**41.33**	**38.91**	**40.01**	**36.08**	**34.60**	**35.32**	**78.48**	**77.98**	**78.79**
DAIMC	48.29	47.46	44.89	44.61	38.45	36.28	83.32	76.83	75.50
DAD+DAIMC	**48.97**	45.06	44.44	43.05	**42.08**	**43.12**	**83.88**	**84.03**	**84.59**
OPIMC	49.24	48.34	44.12	42.98	41.54	35.98	84.89	83.70	80.64
DAD+OPIMC	**57.41**	**54.15**	**56.36**	**49.26**	**45.03**	**46.09**	84.21	**83.93**	**83.64**
EEIMVC	41.02	42.35	40.69	35.34	33.23	28.92	80.62	78.89	76.47
DAD+EEIMVC	**44.11**	**45.02**	**43.08**	**36.55**	**35.92**	**34.38**	**81.41**	**79.92**	**80.03**
PIC	58.82	58.24	56.50	41.73	44.44	43.51	83.99	83.89	83.64
DAD+PIC	**65.70**	**63.74**	**62.74**	**51.94**	**50.29**	**48.66**	**86.07**	**86.28**	**85.09**
IMSR	55.13	38.81	24.60	44.16	27.96	9.12	82.58	74.02	62.75
DAD+IMSR	**69.20**	**69.20**	**69.20**	**58.37**	**58.37**	**58.37**	**88.26**	**88.26**	**88.26**
SRLC	54.21	51.74	48.09	38.99	42.56	32.78	83.22	83.56	80.27
DAD+SRLC	**57.38**	**56.61**	**54.75**	**47.36**	**43.89**	**41.37**	**85.32**	**83.65**	**83.36**
OSLF-IMVC	42.78	38.93	34.91	28.94	24.50	20.63	76.07	74.38	73.53
DAD+OSLF-IMVC	**46.36**	**46.08**	**44.91**	**32.82**	**32.64**	**31.80**	**76.07**	**78.26**	**78.06**
PIMVC	67.50	67.35	66.40	55.95	55.02	52.22	87.30	87.44	86.47
DAD+PIMVC	**67.88**	**68.40**	**66.81**	**56.31**	**55.92**	**53.73**	**87.76**	**88.33**	**87.39**

Figure 7. Ablation experiment on NH-face dataset. (a) Original missing images. (b), (c), and (d) are recovered results via DA-DMVG500, DMVG500, and DMVG1000, respectively.

the-art IMVC methods. Specifically, "DAD+" denotes the processes of filling missing views via DA-DMVG first and then obtaining the clustering result via the corresponding clustering methods. The clustering results of the popular IMVC methods with/without DA-DMVG method on Caltech7 dataset with 0.1, 0.3, and 0.5 missing views are shown in Table 4. We can find that almost all clustering results are further improved by adopting DA-DMVG as a pre-processing step to fill the missing views. For more experimental comparisons and analysis, please refer to Section B in the appendix.

5. Conclusion

In this paper, we propose DMVG, which is the first diffusion-based work for missing view generation. In addition, we also propose a data augmentation strategy to further improve the generation quality of DMVG for missing views, especially for cases with a high missing view rate. Experimental results on the IMVC tasks show that after generating and filling the missing views via our method, better results can be obtained in most cases, especially when the missing view rate is high. However, it is worth noting that filling in missing views may not always improve clustering performance because of the noise introduced by the generated data. Therefore, developing more efficient missing-view generation models that can minimize data noise and designing clustering methods that are robust to noisy data will be important directions for future research.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62372136), the Shenzhen Higher Education Stability Support Program Project (Grant No. GXWD20220811173317002), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515030213), and the Chinese Association for Artificial Intelligence (CAAI)-Huawei MindSpore Open Fund under Grant NoCAAIXSJLJJ-2022-011C.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Cao, X., Zhang, C., Zhou, C., Fu, H., and Foroosh, H. Constrained multi-view video face clustering. *IEEE Transactions on Image Processing*, 24(11):4381–4393, 2015.

Chandrasekaran, G., Nguyen, T. N., and Hemanth D, J. Multimodal sentimental analysis for social media applications: A comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5): e1415, 2021.

Deng, S., Wen, J., Liu, C., Yan, K., Xu, G., and Xu, Y. Projective incomplete multi-view clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 1(1): 1–13, 2023.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 8780–8794, 2021.

Du, G., Zhou, L., Li, Z., Wang, L., and Lü, K. Neighbor-aware deep multi-view clustering via graph convolutional network. *Information Fusion*, 93:330–343, 2023.

Espinosa-Duró, V., Faundez-Zanuy, M., and Mekyska, J. A new face database simultaneously acquired in visible, near-infrared and thermal spectrums. *Cognitive Computation*, 5(1):119–135, 2013.

Greene, D. and Cunningham, P. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the International Conference on Machine Learning*, pp. 377–384, 2006.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1–8, 2021.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 6840–6851, 2020.

Hu, M. and Chen, S. Doubly aligned incomplete multi-view clustering. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 2262–2268, 2018.

Hu, M. and Chen, S. One-pass incomplete multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3838–3845, 2019.

Kang, Z., Lin, Z., Zhu, X., and Xu, W. Structured graph learning for scalable subspace clustering: From single view to multiview. *IEEE Transactions on Cybernetics*, 52 (9):8976–8986, 2021.

Kortli, Y., Jridi, M., Al Falou, A., and Atri, M. Face recognition systems: A survey. *Sensors*, 20(2):342, 2020.

Kumar, A., Rai, P., and Daume, H. Co-regularized multi-view spectral clustering. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 1–9, 2011.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, F.-F., Andreeto, M., Ranzato, M., and Perona, P. Caltech 101, 2022a. URL https://data.caltech.edu/records/20086.

Li, Y., Nie, F., Huang, H., and Huang, J. Large-scale multi-view spectral clustering via bipartite graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2750–2756, 2015.

Li, Z., Tang, C., Zheng, X., Liu, X., Zhang, W., and Zhu, E. High-order correlation preserved incomplete multi-view subspace clustering. *IEEE Transactions on Image Processing*, 31(1):2067–2080, 2022b.

Liang, W., Zhou, S., Xiong, J., Liu, X., Wang, S., Zhu, E., Cai, Z., and Xu, X. Multi-view spectral clustering with high-order optimal neighborhood laplacian matrix. *IEEE Transactions on Knowledge and Data Engineering*, 34 (7):3418–3430, 2020.

Lin, Y., Gou, Y., Liu, Z., Li, B., Lv, J., and Peng, X. Completer: Incomplete multi-view clustering via contrastive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11174–11183, 2021.

Lin, Y., Gou, Y., Liu, X., Bai, J., Lv, J., and Peng, X. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4447–4461, 2023. doi: 10.1109/TPAMI.2022.3197238.

Liu, J., Wang, C., Gao, J., and Han, J. Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of the SIAM International Conference on Data Mining*, pp. 252–260. SIAM, 2013.

Liu, J., Liu, X., Zhang, Y., Zhang, P., Tu, W., Wang, S., Zhou, S., Liang, W., Wang, S., and Yang, Y. Self-representation subspace clustering for incomplete multi-view data. In *Proceedings of the ACM International Conference on Multimedia*, pp. 2726–2734, 2021.

Liu, J., Liu, X., Yang, Y., Liao, Q., and Xia, Y. Contrastive multi-view kernel learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(8):9552–9566, 2023a. doi: 10.1109/TPAMI.2023.3253211. URL https://doi.org/10.1109/TPAMI.2023.3253211.

Liu, M., Liang, K., Hu, D., Yu, H., Liu, Y., Meng, L., Tu, W., Zhou, S., and Liu, X. Tmac: Temporal multi-modal graph learning for acoustic event classification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3365–3374, 2023b.

Liu, M., Liu, Y., LIANG, K., Tu, W., Wang, S., Liu, X., et al. Deep temporal graph clustering. In *The Twelfth International Conference on Learning Representations*, 2023c.

Liu, X., Zhu, X., Li, M., Tang, C., Zhu, E., Yin, J., and Gao, W. Efficient and effective incomplete multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4392–4399, 2019.

Newman, D. Uci machine learning repository. *http://www. ics. uci. edu/˜ mlearn/MLRepository. html*, 2007.

Ren, Z. and Sun, Q. Simultaneous global and local graph structure preserving for multiple kernel clustering. *IEEE transactions on neural networks and learning systems*, 32 (5):1839–1851, 2020.

Ren, Z., Yang, S. X., Sun, Q., and Wang, T. Consensus affinity graph learning for multiple kernel clustering. *IEEE Transactions on Cybernetics*, 51(6):3273–3284, 2020.

Ren, Z., Sun, Q., and Wei, D. Multiple kernel clustering with kernel k-means coupled graph tensor learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 9411–9418, 2021.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Shang, C., Palmer, A., Sun, J., Chen, K.-S., Lu, J., and Bi, J. Vigan: Missing view imputation with generative adversarial networks. In *Proceedings of the IEEE International Conference on Big Data*, pp. 766–775. IEEE, 2017.

Tran, L., Liu, X., Zhou, J., and Jin, R. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1405–1414, 2017.

Wang, H., Zong, L., Liu, B., Yang, Y., and Zhou, W. Spectral perturbation meets incomplete multi-view data. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 3677–3683, 2019.

Wang, Q., Ding, Z., Tao, Z., Gao, Q., and Fu, Y. Generative partial multi-view clustering with adaptive fusion and cycle consistency. *IEEE Transactions on Image Processing*, 30(1):1771–1783, 2021.

Wang, Q., Tao, Z., Gao, Q., and Jiao, L. Multi-view subspace clustering via structured multi-pathway network. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. doi: 10.1109/TNNLS.2022.3213374.

Wang, Q., Tao, Z., Xia, W., Gao, Q., Cao, X., and Jiao, L. Adversarial multiview clustering networks with adaptive fusion. *IEEE transactions on neural networks and learning systems*, 34:7635–7647, 2023. doi: 10.1109/TNNLS.2022.3145048.

Wen, J., Zhang, Z., Xu, Y., Zhang, B., Fei, L., and Liu, H. Unified embedding alignment with missing views inferring for incomplete multi-view clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5393–5400, 2019.

Wen, J., Liu, C., Deng, S., Liu, Y., Fei, L., Yan, K., and Xu, Y. Deep double incomplete multi-view multi-label learning with incomplete labels and missing views. *IEEE Transactions on Neural Networks and Learning Systems*, 2023a.

Wen, J., Zhang, Z., Fei, L., Zhang, B., Xu, Y., Zhang, Z., and Li, J. A survey on incomplete multiview clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(2):1136–1149, 2023b.

Xia, W., Yang, Y., Xue, J.-H., and Wu, B. Towards open-world text-guided face image generation and manipulation. *arXiv preprint arXiv:2104.08910*, 2021.

Xu, C., Liu, H., Guan, Z., Wu, X., Tan, J., and Ling, B. Adversarial incomplete multiview subspace clustering networks. *IEEE Transactions on Cybernetics*, 52(10): 10490–10503, 2021.

Yang, X., Jiaqi, J., Wang, S., Liang, K., Liu, Y., Wen, Y., Liu, S., Zhou, S., Liu, X., and Zhu, E. Dealmvc: Dual contrastive calibration for multi-view clustering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 337–346, 2023.

Zhang, C., Han, Z., Fu, H., Zhou, J. T., Hu, Q., et al. Cpm-nets: cross partial multi-view networks. In *Proceedings of the Advances in Neural Information Processing Systems*, pp. 559–569, 2019.

Zhang, Q., Wei, Y., Han, Z., Fu, H., Peng, X., Deng, C., Hu, Q., Xu, C., Wen, J., Hu, D., et al. Multimodal fusion on

low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*, 2024.

Zhang, Y., Liu, X., Wang, S., Liu, J., Dai, S., and Zhu, E. One-stage incomplete multi-view clustering via late fusion. In *Proceedings of the ACM International Conference on Multimedia*, pp. 2717–2725, 2021.

Zhao, H., Liu, H., and Fu, Y. Incomplete multi-modal visual data grouping. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 2392–2398, 2016.

Zhou, P. and Du, L. Learnable graph filter for multi-view clustering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3089–3098, 2023.

Zhou, P., Du, L., and Li, X. Adaptive consensus clustering for multiple k-means via base results refining. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Zhou, S., Liu, X., Liu, J., Guo, X., Zhao, Y., Zhu, E., Zhai, Y., Yin, J., and Gao, W. Multi-view spectral clustering with optimal neighborhood laplacian matrix. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6965–6972, 2020.

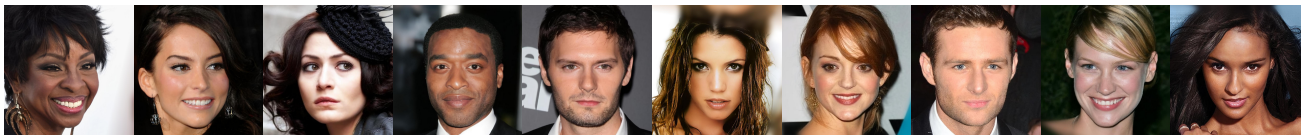
Zhu, X., Zhang, S., Zhu, Y., Zheng, W., and Yang, Y. Self-weighted multi-view fuzzy clustering. *ACM transactions on knowledge discovery from data*, 14(4):1–17, 2020.

Zhuge, W., Hou, C., Liu, X., Tao, H., and Yi, D. Simultaneous representation learning and clustering for incomplete multi-view data. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 4482–4488, 2019.

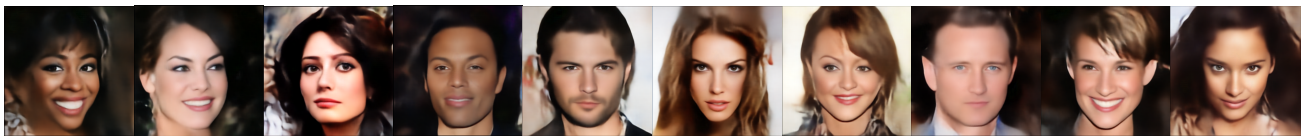
A. More Experiments on Missing-view Generation

Firstly, we show more generated examples on Multi-Modal CelebA-HQ in Figures 8 and 9.

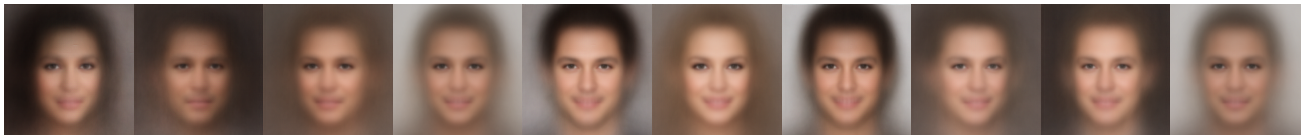
(a) Original sketch images



(b) Original RGB missing images (not used in model training)



(c) Generated RGB images via DMVG



(d) Generated RGB images via VIGAN

Figure 8. Generating RGB missing images according to their sketch images on Multi-Modal CelebA-HQ.

(a) Original RGB images

(b) Original sketch missing images (not used in model training)

(c) Generated sketch images via DMVG

(d) Generated sketch images via VIGAN

Figure 9. Generating sketch missing images according to their RGB images on Multi-Modal CelebA-HQ.

Secondly, we compare our proposed method with VIGAN and CRA on several datasets usually used in IMVC, *i.e.*, BBCSports (Greene & Cunningham, 2006), Handwritten (Newman, 2007), Caltech7 (Li et al., 2015), and Animal (Zhang

et al., 2019). Because BBCSports, Handwritten, and Caltech7 have more than two views, we compare DA-DMVG with CRA on these datasets with 0.1, 0.3, and 0.5 missing views. The quantitative results in terms of RMSE, NMSE, and PSNR are shown in Tables 5, 6, and 7, respectively. It is obvious that DA-DMVG is always better than CRA. Furthermore, Figure 10 shows many visualization results on Handwritten with 0.1 missing views. DA-DMVG nearly recovered the missing view 1 perfectly.

Table 5. RMSE, NMSE, and PSNR of CRA and DA-DMVG on BBCSport with 10%, 30%, and 50% missing views.

		RMSE↓			NMSE↓			PSNR↑		
VIEW	METHOD	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
2*1	CRA	0.06	0.06	0.06	2.33	3.89	4.87	24.90	25.03	24.91
	DA-DMVG	**0.05**	**0.03**	**0.03**	**1.60**	**1.42**	**1.53**	**26.54**	**29.41**	**29.94**
2*2	CRA	0.06	0.06	0.05	4.21	4.74	3.88	24.79	24.93	25.96
	DA-DMVG	**0.03**	**0.03**	**0.04**	**1.16**	**1.26**	**1.99**	**30.39**	**30.70**	**28.85**
2*3	CRA	0.05	0.06	0.05	3.52	6.71	5.04	26.11	24.75	26.16
	DA-DMVG	**0.03**	**0.04**	**0.03**	**1.46**	**2.49**	**1.65**	**29.92**	**29.05**	**31.01**
2*4	CRA	0.06	0.05	0.06	2.17	10.16	11.74	24.59	25.59	24.89
	DA-DMVG	**0.04**	**0.02**	**0.02**	**1.10**	**2.30**	**1.85**	**27.51**	**32.05**	**32.91**

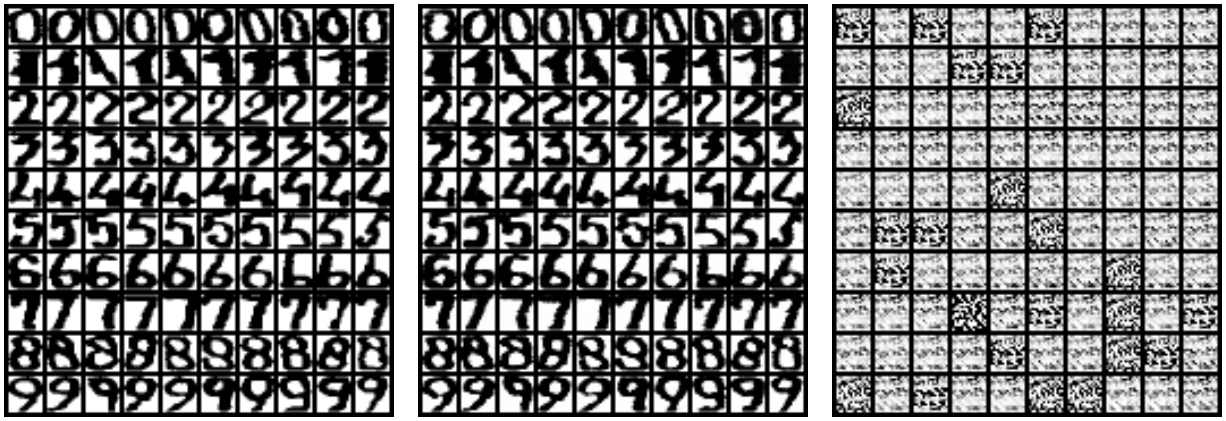
Table 6. RMSE, NMSE, and PSNR of CRA and DA-DMVG on Handwritten with 10%, 30%, and 50% missing views.

		RMSE↓			NMSE↓			PSNR↑		
VIEW	METHOD	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
2*1	CRA	0.45	0.46	0.50	1.00	1.02	1.22	6.86	6.78	5.99
	DA-DMVG	**0.25**	**0.22**	**0.29**	**0.31**	**0.22**	**0.41**	**11.91**	**13.25**	**10.72**
2*2	CRA	0.35	0.30	0.38	6.43	5.31	9.21	9.17	10.41	8.30
	DA-DMVG	**0.08**	**0.08**	**0.09**	**0.37**	**0.40**	**0.52**	**21.56**	**21.64**	**20.74**
2*3	CRA	0.06	0.06	0.06	0.05	0.05	0.05	24.22	24.90	24.91
	DA-DMVG	**0.02**	**0.02**	**0.05**	**0.00**	**0.01**	**0.03**	**35.27**	**32.36**	**26.62**
2*4	CRA	0.10	0.10	0.09	0.39	0.41	0.39	19.86	19.82	20.55
	DA-DMVG	**0.02**	**0.03**	**0.04**	**0.02**	**0.04**	**0.09**	**33.07**	**30.47**	**27.14**
2*5	CRA	0.11	0.15	0.17	2.07	3.61	4.25	18.80	16.48	15.56
	DA-DMVG	**0.03**	**0.04**	**0.06**	**0.13**	**0.26**	**0.54**	**30.67**	**27.85**	**24.53**

Table 7. RMSE, NMSE, and PSNR of CRA and DA-DMVG on Caltech7 with 10%, 30%, and 50% missing views.

		RMSE↓			NMSE↓			PSNR↑		
VIEW	METHOD	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
2*1	CRA	0.20	0.22	0.14	9.90	10.82	6.00	13.80	13.15	17.22
	DA-DMVG	**0.05**	**0.06**	**0.03**	**0.62**	**0.89**	**0.34**	**25.39**	**23.99**	**29.67**
2*2	CRA	0.19	0.20	0.23	0.46	0.49	0.67	14.33	13.94	12.70
	DA-DMVG	**0.14**	**0.13**	**0.11**	**0.24**	**0.20**	**0.16**	**17.07**	**17.93**	**18.83**
2*3	CRA	0.10	0.13	0.14	5.05	9.59	10.78	20.02	17.80	17.36
	DA-DMVG	**0.03**	**0.03**	**0.02**	**0.45**	**0.21**	**0.16**	**30.56**	**34.46**	**35.74**
2*4	CRA	0.28	0.31	0.31	1.01	1.18	1.17	10.93	10.24	10.28
	DA-DMVG	**0.20**	**0.21**	**0.21**	**0.50**	**0.55**	**0.58**	**13.93**	**13.51**	**13.35**
2*5	CRA	0.25	0.34	0.34	5.38	12.40	11.92	12.16	9.26	9.48
	DA-DMVG	**0.08**	**0.06**	**0.07**	**0.54**	**0.42**	**0.52**	**22.15**	**23.93**	**23.07**
2*6	CRA	0.35	0.43	0.28	15.99	25.80	10.84	9.04	7.28	11.01
	DA-DMVG	**0.04**	**0.08**	**0.06**	**0.24**	**0.79**	**0.49**	**27.20**	**22.40**	**24.44**

Because Animal is a dual-view dataset, we cannot augment more data via the proposed data augmentation strategy. So we compare DMVG with VIGAN on this dataset with 0.3, 0.5, and 0.7 paired views. The quantitative results in terms of RMSE, NMSE, and PSNR are shown in Table 8, which demonstrate that DMVG outperforms VIGAN all the time.



(a) Original missing view 1 (b) Recovered view 1 via DA-DMVG (c) Recovered view 1 via CRA

Figure 10. Generating missing view 1 according to other available views on Handwritten dataset with 10% missing views.

Table 8. RMSE, NMSE, and PSNR of VIGAN and DMVG on Animal with 30%, 50%, and 70% paired views.

		RMSE↓			NMSE↓			PSNR↑		
VIEW	METHOD	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7
2*1	VIGAN	0.08	**0.07**	0.07	1.55	1.24	1.28	21.99	22.83	23.49
	DMVG	**0.07**	**0.07**	**0.06**	**1.15**	**1.18**	**1.15**	**23.28**	**23.08**	**23.93**
2*2	VIGAN	0.08	**0.08**	0.09	1.20	1.22	1.20	21.88	21.70	21.31
	DMVG	**0.07**	**0.08**	**0.08**	**0.88**	**1.16**	**1.01**	**23.22**	**21.92**	**22.06**

B. More Experiments on IMVC after Missing-view Generation

Here are more experiments on IMVC after missing-view generation. Specifically, we filled BBCSport and Handwritten via DA-DMVG, and Animal via DMVG (since Animal is a dual-view dataset and cannot undergo data augmentation). Specifically,"D"/"DAD+" denotes the processes of filling missing views via DMVG/DA-DMVG first and then clustering via the corresponding clustering methods. The clustering results of the popular IMVC methods with/without our proposed DMVG/DA-DMVG methods on BBCSports, Handwritten, and Animal are shown in Tables 9, 10, and 11, respectively. Combined with the results on Caltech7 shown in Table 4, we obtain the following conclusions:

(1) DMVG and DA-DMVG generally enhance clustering performance while are less effective for datasets with fewer samples. In all experiments, DMVG or DA-DMVG improved incomplete multi-view clustering performance in about 76% cases, indicating the great potential of missing-view generation for IMVC. Specifically, clustering performance improved by about 46% on BBCSport, 70% on Handwritten, 96% on Caltech7, and 88% on Animal, respectively. However, experiments on BBCSport suggest that missing-view generation does not always yield benefits, because we just use a small number of samples on BBCSport, which may hinder DA-DMVG from effectively learning the intrinsic correlations from conditional views to the target view.

(2) Empirical filling methods like zero-filling and mean-filling negatively impact clustering. Except experiments on BBCSport, all experiments combining the missing-view generation with MVC algorithms show improved clustering performance, because these MVC algorithms have to rely on zero-filling or mean-filling for IMVC without DMVG or DA-DMVG integration. Zero-filling and mean-filling tend to cluster samples with missing-views together, significantly affecting clustering results. In contrast, DMVG and DA-DMVG accurately predict missing views, often significantly improving performance when combined with methods like BSV, where ACC improves by exceeding 10% across different datasets and missing rates.

(3) Existing IMVC methods often fail to fully exploit the information contained in available views. Except BBCSport, about 75% experiments combining missing-view generation with IMVC methods show improved clustering performance. Although IMVC methods can independently perform IMVC, DMVG or DA-DMVG often further improve their clustering performance. This indicates that existing IMVC methods do not fully exploit information from available views, that is, focusing on consistent information across views while neglecting complementary information. On the contrary, DMVG and

Table 9. ACC(%) and NMI(%) of different methods on BBCSport with 10%, 30%, 50%, and 70% missing views.

	ACC				NMI			
METHOD	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7
BSV	58.62	51.31	44.03	36.43	43.73	31.03	21.40	12.05
DAD+BSV	**62.41**	**62.59**	**53.97**	**39.31**	42.22	**40.24**	**29.94**	**12.67**
CONCAT	70.62	58.72	33.21	35.95	61.69	38.92	18.61	8.58
DAD+CONCAT	62.41	**63.97**	**49.83**	33.28	49.99	**48.88**	**29.42**	8.29
MULTINMF	48.58	42.75	40.34	35.69	23.48	18.25	14.79	7.84
DAD+MULTINMF	45.86	**46.67**	**42.24**	**36.03**	21.09	**20.50**	**15.92**	**8.67**
CCR-MVSC	72.76	70.06	61.38	35.86	62.79	57.69	39.55	14.11
DAD+CCR-MVSC	72.41	65.69	60.00	**36.55**	**63.54**	52.71	37.40	13.34
DAIMC	68.62	63.45	56.89	39.59	56.62	50.17	37.89	17.16
DAD+DAIMC	**71.03**	**66.55**	52.93	35.00	**57.49**	**51.49**	32.34	12.55
OPIMC	54.14	52.93	45.69	44.34	35.66	31.56	21.75	14.65
DAD+OPIMC	35.17	43.10	36.55	33.10	11.15	20.68	13.18	4.51
EEIMVC	76.03	73.45	62.76	47.41	65.34	61.25	46.91	25.95
DAD+EEIMVC	**76.90**	67.24	56.90	34.48	**67.26**	53.92	33.36	9.56
PIC	75.52	74.48	69.48	31.89	70.94	64.18	53.91	9.99
DAD+PIC	75.34	72.59	68.62	**39.31**	68.89	**65.61**	53.73	**16.28**
IMSR	78.45	72.41	63.45	41.21	69.39	61.56	43.35	20.00
DAD+IMSR	**84.31**	**84.31**	**84.14**	**84.43**	**72.02**	**72.02**	**72.79**	**72.85**
SRLC	69.83	57.24	43.28	34.83	50.76	35.96	21.37	9.12
DAD+SRLC	66.03	**61.21**	**59.66**	**37.93**	**52.54**	**42.19**	**35.36**	**11.88**
OSLF-IMVC	75.86	75.34	61.21	45.52	67.17	67.30	44.73	21.47
DAD+OSLF-IMVC	**77.59**	73.45	**63.10**	37.24	**68.36**	60.71	40.98	12.41
PIMVC	79.66	75.17	73.97	52.07	71.12	64.22	60.81	29.32
DAD+PIMVC	79.66	71.55	67.07	39.14	70.86	58.71	49.07	15.64

Table 10. ACC(%) and NMI(%) of different methods on Handwritten with 10%, 30%, 50%, and 70% missing views.

	ACC				NMI			
METHOD	0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7
BSV	68.27	51.49	38.24	27.15	62.82	47.01	32.21	19.48
DAD+BSV	**79.89**	**82.08**	**76.15**	**66.09**	**75.18**	**74.83**	**67.05**	**51.72**
CONCAT	75.06	55.48	42.19	28.31	73.08	51.66	38.24	23.50
DAD+CONCAT	**88.73**	**89.56**	**85.62**	**67.38**	**81.97**	**81.34**	**75.61**	**53.10**
MULTINMF	82.35	71.74	52.03	31.85	72.05	60.11	41.99	20.88
DAD+MULTINMF	**82.46**	**80.40**	**81.12**	**69.00**	**72.93**	**71.32**	**69.55**	**56.45**
CCR-MVSC	74.61	73.17	70.15	64.62	70.90	68.23	62.86	53.14
DAD+CCR-MVSC	**75.92**	**76.15**	**74.46**	**73.09**	**71.51**	**71.18**	**67.70**	**60.44**
DAIMC	88.86	86.73	81.92	60.44	79.78	76.65	68.77	47.10
DAD+DAIMC	85.81	82.36	80.57	**70.69**	76.82	74.26	70.22	**54.17**
OPIMC	80.20	76.45	69.50	56.66	77.26	73.74	66.57	51.86
DAD+OPIMC	75.63	72.92	**72.62**	**63.06**	73.53	70.54	**67.41**	51.39
EEIMVC	88.60	85.23	76.70	51.74	78.64	73.30	62.26	40.21
DAD+EEIMVC	86.12	83.22	**84.19**	**77.71**	76.82	**74.19**	**72.90**	**62.90**
PIC	84.20	83.90	83.24	80.97	85.41	84.79	82.25	77.56
DAD+PIC	83.69	83.25	80.76	75.17	**86.16**	84.77	82.04	70.26
IMSR	90.36	89.74	83.68	62.10	83.26	81.57	72.81	53.92
DAD+IMSR	87.03	86.98	**86.99**	**87.05**	79.79	79.77	**79.75**	**79.81**
SRLC	95.09	88.62	81.04	69.46	90.15	84.27	75.33	62.70
DAD+SRLC	**96.06**	**94.84**	**87.62**	**81.10**	**91.66**	**89.27**	**81.14**	**69.00**
OSLF-IMVC	75.04	70.21	55.17	35.79	67.16	60.98	44.70	27.07
DAD+OSLF-IMVC	**79.60**	**81.46**	**80.66**	**69.33**	**74.68**	**73.58**	**69.92**	**57.99**
PIMVC	94.88	93.79	91.23	88.61	89.74	87.71	83.88	79.56
DAD+PIMVC	**94.97**	**93.99**	**91.27**	79.51	89.25	**87.92**	83.25	67.96

DA-DMVG utilize all information from available views to predict missing views as accurately as possible.

Table 11. ACC(%), NMI(%), and PUR(%) of different methods on Animal with 30%, 50%, and 70% paired views.

	ACC			NMI			PUR		
METHOD	0.3	0.5	0.7	0.3	0.5	0.7	0.3	0.5	0.7
BSV	42.05	48.63	56.22	48.16	55.91	63.99	45.20	52.26	60.31
D+BSV	**55.75**	**60.49**	**64.95**	**63.23**	**67.16**	**71.73**	**61.20**	**65.42**	**70.14**
CONCAT	42.79	49.34	53.99	55.46	59.31	63.88	48.12	53.24	59.26
D+CONCAT	**52.91**	**58.02**	**59.63**	**60.06**	**64.75**	**68.25**	**58.11**	**62.78**	**65.46**
CCR-MVSC	52.03	54.72	56.73	55.31	58.31	61.71	56.28	59.36	61.98
D+CCR-MVSC	**54.17**	**57.83**	**58.29**	**57.48**	**60.87**	**63.43**	**58.25**	**61.94**	**63.33**
DAIMC	50.18	53.87	56.42	55.03	59.36	62.76	54.82	59.51	62.12
D+DAIMC	**53.75**	**58.27**	**59.25**	**59.60**	**64.16**	**67.14**	**58.75**	**63.41**	**65.62**
OPIMC	46.33	53.14	53.88	52.34	58.51	62.04	49.49	56.23	57.91
D+OPIMC	**53.58**	**57.70**	**58.52**	**60.32**	**64.40**	**67.77**	**58.43**	**62.15**	**64.42**
EEIMVC	45.90	53.34	57.15	53.72	57.92	62.02	51.30	57.40	61.74
D+EEIMVC	**56.07**	**59.41**	**61.15**	**59.94**	**63.40**	**66.37**	**60.57**	**63.21**	**65.61**
PIC	55.94	56.84	57.67	62.35	64.37	65.82	63.07	64.75	65.42
D+PIC	**56.01**	51.48	55.36	**62.71**	60.75	65.51	**63.64**	60.81	65.30
IMSR	47.02	53.15	58.38	55.87	60.00	65.43	52.61	57.80	63.78
D+IMSR	**59.60**	**60.03**	**59.72**	**68.02**	**68.10**	**67.44**	**65.63**	**65.75**	**65.47**
SRLC	51.14	53.93	55.76	56.77	60.43	63.54	58.08	61.24	63.90
D+SRLC	**51.42**	**54.17**	**57.87**	**57.14**	**61.16**	**65.12**	**58.26**	**62.18**	**65.48**
OSLF-IMVC	40.53	48.24	55.07	50.53	54.23	58.51	45.31	51.68	56.93
D+OSLF-IMVC	**54.64**	**58.19**	**57.99**	**57.23**	**61.00**	**62.91**	**56.82**	**59.98**	**60.36**
PIMVC	55.56	57.47	59.24	61.54	63.92	65.84	60.45	63.04	64.75
D+PIMVC	54.09	**57.54**	**59.28**	60.23	62.61	**65.88**	59.31	61.83	64.44

(4) Missing-view generation significantly reduces sensitivity to missing view rates in existing methods. After combining DMVG or DA-DMVG with existing clustering methods, models showed much lower sensitivity to missing view rates.

(5) Noise introduced by missing view generation can negatively impact clustering. We also observed that in some experiments (except experiments on BBCSport), DMVG and DA-DMVG not only failed to enhance clustering performance, but even degraded clustering performance, such as the experimental results of DAIMC and PIC on Handwritten. It is speculated that the impact of missing-view generation for IMVC is two-sided. On one hand, missing-view generation essentially learns correlations between different views, aiding clustering models in easily learning consistent representations shared across views. On the other hand, noise introduced by DMVG or DA-DMVG can reduce clustering performance. When the disadvantages of noise outweigh the benefits of missing-view generation, clustering performance may decline.