Realize Generative yet Complete Latent Representation for Incomplete Multi-view Learning

Hongmin Cai[1], *Senior Member, IEEE*, Weitian Huang[1], Sirui Yang[1], Siqi Ding[1], Yue Zhang[2], Bin Hu[3], *Fellow, IEEE*, Fa Zhang[3], Yiu-ming Cheung[4,*], *Fellow, IEEE*

Abstract—In multi-view environment, it would yield missing observations due to the limitation of the observation process. The most current representation learning methods struggle to explore complete information by lacking either cross-generative via simply filling in missing view data, or solidative via inferring a consistent representation among the existing views. To address this problem, we propose a deep generative model to learn a complete generative latent representation, namely Complete Multi-view Variational Auto-Encoders (CMVAE), which models the generation of the multiple views from a complete latent variable represented by a mixture of Gaussian distributions. Thus, the missing view can be fully characterized by the latent variables and is resolved by estimating its posterior distribution. Accordingly, a novel variational lower bound is introduced to integrate view-invariant information into posterior inference to enhance the solidative of the learned latent representation. The intrinsic correlations between views are mined to seek cross-view generality, and information leading to missing views is fused by view weights to reach solidity. Benchmark experimental results in clustering, classification, and cross-view image generation tasks demonstrate the superiority of CMVAE, while time complexity and parameter sensitivity analyses illustrate the efficiency and robustness. Additionally, application to bioinformatic data exemplifies its practical significance.

Index Terms—Multi-view learning, incomplete multi-view problem, representation learning, deep generative models.

1 Introduction

Multiple views of data in real-world applications are collected by different measurements to represent various characteristics of an object. Concrete examples include reporting news in different languages, describing events with images, audio and text, and detecting organs through different imaging mechanisms to obtain multi-modal medical images. These semantically coherent multi-view samples are connected by a consensus representation. Typically, limitations or deviations in measurement methods result in individual views containing insufficient information, while different views can complement each other [1]. In several radical cases, there may even be missing views, making it difficult to obtain complementary information from other views, so-called *incomplete multi-view problem*.

In recent years, there has been a growing body of research in multi-view representation learning, which is concerned with the problem of exploiting complementary information to learn integrated representations. Representative techniques include correlation-based [2], [3], similarity-based [4], [5], graphical model-based [6], [7] and neural network-based [8], [9] multi-view learning approaches. They perform model learning based on the assumption that all views of each sample are fully observed. However, for incomplete multi-view data, these methods will inevitably degrade or even collapse.

To solve the incomplete multi-view problem, various incomplete multi-view learning algorithms have been proposed recently. In concrete, the prevalent solutions can be roughly classified into three strategies. The first strategy attempts to fill in the missing view data by assuming that the learned multi-view representation contains complete information, followed by off-the-shelf task driving processing. [10], [11] presented the approaches to first impute the missing values by matrix completion, and then performed weighted non-negative matrix factorization (NMF) by setting the filled data lower weight. Besides, generative adversarial networks (GAN) based approaches [12], [13] leveraged the power of generative network to impute the missing views. However, this strategy relies heavily on the effectiveness of data completion and is usually ineffective when the missing rate is high. The second strategy is to group samples into multiple paired subsets according to the availability of data sources by assuming the data for each subset is complete and then learn multiple models on these groupings for post-fusion [14], [15]. Although it is more efficient than learning on each individual view, this inflexible grouping strategy greatly increases computational complexity when facing data with a large number of views. The last strategy focuses on inferring the latent information on the missing views by assuming that the cross-

[1]*School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong, China.*
[2]*School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, Guangdong, China.*
[3]*School of Medical Technology, Beijing Institute of Technology, Beijing, China.*
[4]*Department of Computer Science, Hong Kong Baptist University, Hong Kong.*
**Yiu-ming Cheung is the Corresponding Author, whose E-mail is: ymc@comp.hkbu.edu.hk.*

views are generative. Partial multi-view clustering [16], [17] presented a joint partial alignment method to explore the complementary and consensus information of the available views. In [18], [19], the embedding space of both views was learned by cross-view learning, followed by distilling latent information across views. This allows for more efficient usage of available views to infer potentially shared information, but the completeness of the latent representation is not guaranteed. To deal with general requirements, we summarize the following desiderata for incomplete multi-view representation learning:

1) **Completeness.** The learned multi-view representation contains complete information, which describes task-relevant information for all different views of observation.
2) **Cross-generative.** Multi-view representation learned from available views has the ability to generate missing samples, which can also corroborate the completeness of the learned representation.
3) **Solidative.** There exists conservative information inherent to multi-view sampling that is not altered by the absence of views, e.g., the weight of views, the intrinsic correlations between views. Making full use of this information can enhance the solidative of the learned representation.

In this paper, we propose a deep generative model for multi-view representation learning, namely Complete Multi-view Variational Auto-Encoders (CMVAE). As demonstrated in Fig. 1, multi-view observations are considered to be generated from a shared latent variable containing various attributes, which is therefore assumed to be a mixture of densities from multiple variational posteriors. For incomplete multi-view problem, CMVAE seeks to learn a complete generative latent representation accompanied by view-invariant information. First, view-peculiar latent variables are inferred from the corresponding posteriors. Then, the implicit information of the missing views can be predicted from the existing views by the intrinsic correlations between them. By weighting the integration of multiple view-peculiar latent variables, a complete generative latent variable is finally obtained.

The main contributions of our work are as follows,

- A deep multi-view generative model is proposed to learn **complete** generative latent representations. By using variational inference, a novel variational lower bound of joint likelihood is developed with the introduction of joint variational posterior as a Gaussian mixture.
- By optimizing the posterior for each view, a shared Gaussian mixture distribution variable can be inferred, from which multi-view variants can be derived to **cross-generate** data samples from another view.
- View-invariant information is learned to enhance the **solidative** of the latent representation when encountering view missing. Precisely, the implicit information of missing views can be predicted by modelling the linear transformations between view-peculiar latent variables, resulting in the view weights being invariant to ensure the accurate depiction of mixture Gaussian distributions.
- From both quantitative and qualitative perspectives, the results of the benchmark experiments demonstrate the superiority of the learned generative yet complete latent representation, while the time complexity and parameter sensitivity analyses illustrate the efficiency and robustness of CMVAE. In addition, the application on bioinformatics data verifies its the practical value.

2 Related Works

2.1 Multi-view Representation Learning

Multi-view representation learning is defined as a representation learning procedure for discovering the underlying patterns of multi-view data. Canonical Correlation Analysis (CCA) [20] and its variants [2], [3], [21], [22] are typical representation learning model for two views data. CCA-based methods aim to learn the consistent representations by maximizing the total correlation. Recently, many notable multi-view representation learning approaches have been proposed to handle more than two views. MDcR [23] applies kernel matching to regularize the correlation between multiple views in the common low-dimensional latent space. DMF-MVC [24] extracts consistent representation by leveraging semi-non-negative matrix factorization. MULPP [25] utilizes pairwise and higher-order correlations to realize flexible view consistency while maintaining local structure to obtain complementary information. To combine information specific to each single view, SIMM [26] introduces confusion adversarial loss and orthogonality constraints to exploit view-shared and view-specific representations. DGF [27] explicitly models both multi-view consistency as well as multi-view inconsistency in a unified optimization model. In addition, there are many outstanding works leveraging probabilistic generative models for representation learning, which will be introduced in detail in Section 2.3. Intuitively, the consistency information between views decreases with the number of views, while the inconsistency is opposite, making multi-view consensus learning methods difficult to handle. This paper focuses more on extracting complete information from multiple views to preserve the characteristics of different views.

2.2 Incomplete Multi-view Representation Learning

To deal with the incomplete multi-view problem, an increasing number of studies have recently focused on incomplete multi-view representation learning. They can be classified into three categories based on the exploitation of cross-view information. (1) **Missing filling** method aims at imputing the missing data to form the complete multi-view data and then utilizing conventional multi-view learning technique. CoKL [28] collectively completes the kernel matrices of the datasets by optimizing the alignment of common instances. MVL-IV [29] generates incomplete view data from the shared subspace learned from the observed view. CRA [30] stacks residual autoencoders to learn complex relationship among multiple views to impute the missing data. VIGAN [12] utilizes GAN to generate the missing

views and then learns the shared latent space of all views. (2) **Grouping-and-learning** method is to group samples into multiple paired subsets based on available data, which are then divided into multiple learning tasks. iMSF [14] divides samples based on the availability of data sources, and then uses sparse learning to learn shared feature sets. IMG [31] separates the samples into complete and incomplete multi-view data. A compact global structure is then learned by using Laplacian graphs of complete instances in a low-dimensional space. MoPoE-VAE [15] trains different multi-view models by different subsets of groupings to cope with different view missing cases. (3) **Cross-view learning** method infers potential information on the missing data by performing cross-view learning directly from the existing view data. PVC [16] projects the incomplete data into a low-dimensional common subspace regularized by l_F-norm and l_1-norm. iCmSC [19] explores correlations between incomplete cross-view data and learns a consistent subspace representation to improve clustering performance. CPM-nets [17] partitions partial multi-view data by focusing on the completeness and generality of the learned representations. GSRIMC [32] focuses on exploring the useful information behind the subgraph structure, while avoiding to perform complex feature recovery tasks. DCP [33] unifies consistent representation learning and missing data recovery by jointly optimizing dual constraint loss and dual prediction loss from an information-theoretic perspective. This paper is concerned with this category, i.e., making full use of existing data to predict the hidden features of missing views and integrate them into a unified representation.

2.3 VAE in Multi-view Scheme

We begin with a natural assumption that there is a shared latent variable on multiple observations generated from multi-view measurment [34]. Once the variant samples are well generated together, the completeness of the latent representation is achieved. However, the intractable integral calculations involved make it difficult to optimize this generative model directly, i.e., maximizing the joint likelihood function. Variational inference technique is one of the solutions used to convert the difficult computation problem into optimization problems. Variational Auto-encoders (VAE) [35] is a successful paradigm on single-view data by combining the deep neural network under the framework of Stochastic Gradient Variational Bayes (SGVB).

Recently, many generative models leveraging the VAE framework have been proposed for multi-view representation learning. We review the models for constructing variational lower bounds by introducing different joint variational posteriors that elaborate specific inference processes for latent representations. VCCAP [3] utilizes dual variational autoencoders to nonlinearly project two view data into a consistent latent space by maximizing the correlation between views. In DMVC-VAE [40], an auto-weighted fusion module is embedded into the posterior inference process to obtain the shared latent representation which is set as a mixture of Gaussian distributions for clustering. MVAE [36] models the joint posterior as a Product-of-Experts (PoE) [37]. PoE produces a clearer distribution by aggregating information from multiple unimodal posteriors, but this is not conducive to optimizing the individual posteriors, which is important for learning a balanced distribution. MMVAE [38] assumes that the joint posterior is a Mixture-of-Experts (MoE). MMVAE only takes unimodal posteriors into account during training by pairwise optimization, such that the latent representation of any view can reconstruct observations from other views and its own view. The drawback is that information from other posteriors cannot be combined in one-pass inference. mmJSD [39], on the other hand, adds learning of a common latent variable to aggregate view-peculiar latent variables. MoPoE-VAE [15] combines PoE and MoE to construct the variational lower bound of joint likelihood, theoretically combining the advantages of each. They are computationally scalable to the number of views, it decomposes all m view data into 2^m subsets, and optimizes the 2^m combined encoders separately to deal with any missing view. However, this would render the method intractable to handle for tasks with a large number of views. The aforementioned incomplete multi-view learning methods either suffer from high computational complexity or ignore the difference information between views, which motivates the proposed CMVAE to reduce the computational complexity of multi-view VAEs and learn generative yet complete representations to improve the performance of clustering or classification.

3 Proposed Method

In this section, a generic multi-view probabilistic generative model, the vanilla multi-view VAE (VMVAE), is first proposed for multi-view representation learning. To efficiently optimize the latent variable model, a variational lower bound for the joint likelihood function is constructed by introducing a mixture of Gaussian distributions as the joint variational posterior. Next, to meet the three desiderata for incomplete multi-view representation learning, we present CMVAE where a novel joint variational posterior is proposed, which extract implicit correlations among views and learns invariant weights across views to realize a complete generative latent representation.

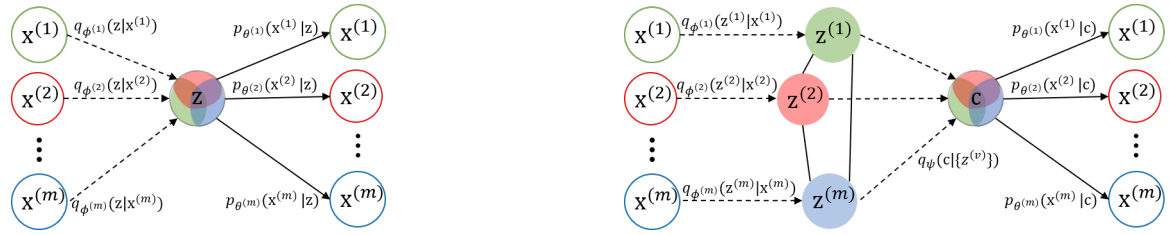
3.1 Vanilla Multi-view VAE

Given a m views dataset $\boldsymbol{\mathcal{X}} = \{\mathbf{X}^{(v)} \in \mathbb{R}^{n\times d_v}\}_{v=1}^{m}$, each view collection consisting of n i.i.d samples, $\mathbf{X}^{(v)} = \{\mathbf{x}_1^{(v)}, \mathbf{x}_2^{(v)}, ..., \mathbf{x}_n^{(v)}\}$. We denote uniform sampling from this finite dataset as $\widetilde{p}(\boldsymbol{\mathcal{X}})$. To estimate the true density of multi-view variables, one aims to approximate the real distribution $p(\boldsymbol{\mathcal{X}}) \in \mathcal{P}$ from the hypothesis space of distribution family $\mathcal{P}$ by $\widetilde{p}(\boldsymbol{\mathcal{X}})$ nicely via minimization their KL divergence,

$$\begin{aligned}\min D_{\mathrm{KL}}\left(\widetilde{p}(\boldsymbol{\mathcal{X}})\|p(\boldsymbol{\mathcal{X}})\right) &= \max \mathbb{E}_{\widetilde{p}(\boldsymbol{\mathcal{X}})}\left[\log p(\boldsymbol{\mathcal{X}})\right] \\ &= \max \frac{1}{n}\sum_{i=1}^{n} \log p(\{\mathbf{x}_i^{(v)}\}_{v=1}^{m}), \quad (1)\end{aligned}$$

Note that minimizing the KL divergence is equivalent to maximizing the log likelihood. To ease the reading, we will omit the subscript i and denote $\{*^{(v)}\}_{v=1}^{m}$ as $\{*^{(v)}\}$ in the following.

Under the assumption that the density $p(\{\mathbf{x}^{(v)}\})$ is achieved through the marginalization of a shared latent



(a) Vanilla Multi-view VAE (b) Complete Multi-view VAE

Fig. 1. The solid arrows represent the process of generation, the solid lines indicate the intrinsic correlations between views, while the dashed arrows represent the process of variational inference. (The sampling operation is omitted). (a) Multi-view observations are obtained from a shared latent variable through different generative processes. The latent variable is assumed to represent a mixture of multiple posteriors mixing different semantic properties. (b) Complete multi-view VAE is proposed to model the intrinsic transformations between views and preserve the weight of views, which can facilitate the synthesis of the complete latent variable.

continuous variable $\mathbf{z}$, the generation process of multi-view variables can be formulated by,

$$p(\{\mathbf{x}^{(v)}\}) = \int \prod_{v=1}^{m} p_{\theta^{(v)}}(\mathbf{x}^{(v)}|\mathbf{z})p(\mathbf{z})d\mathbf{z}. \tag{2}$$

Since the integral is intractable, it is potentially difficult to directly calculate the marginalization. Similar to VAE [35], we turn to maximize the Lower BOund on the Evidence (ELBO) $\mathcal{L}_{\text{ELBO}}(p(\{\mathbf{x}^{(v)}\}))$ by introducing a joint variational posterior $q(\mathbf{z}|\{\mathbf{x}^{(v)}\})$,

$$\begin{aligned}\log p(\{\mathbf{x}^{(v)}\}) \geq & \mathcal{L}_{\text{ELBO}}(\{\mathbf{x}^{(v)}\}) \\ = & - D_{\text{KL}}\left(q(\mathbf{z}|\{\mathbf{x}^{(v)}\})\|p(\mathbf{z})\right) \\ & + \mathbb{E}_{q(\mathbf{z}|\{\mathbf{x}^{(v)}\})}\left[\log p(\{\mathbf{x}^{(v)}\}|\mathbf{z})\right].\end{aligned} \tag{3}$$

The former is KL divergence from the prior $p(\mathbf{z})$ to the joint variational posterior $q(\mathbf{z}|\{\mathbf{x}^{(v)}\})$, which drives variational inference close to our hypothesis on $\mathbf{z}$. The latter reveals the process of variational inference and generation.

Once a specific posterior model is established, a shared latent representation can be learned from multiple variables through the inference process. Therefore, it is crucial to choose a highly expressive and easily computable density as the joint variational posterior. Different strategies for modelling the joint posterior inference process are discussed in Section 2.3. We believe that the complete implicit semantic information of events is complexly distributed, where only limited properties are actually observed at once. Therefore, we prefer to model the joint variational posterior with a mixture of Gaussian distributions,

$$\begin{aligned}q(\mathbf{z}|\{\mathbf{x}^{(v)}\}) &= \sum_{v=1}^{m} \lambda_v q_{\phi^{(v)}}(\mathbf{z}|\mathbf{x}^{(v)}) \\ &= \sum_{v=1}^{m} \lambda_v \mathcal{N}(\mathbf{z};\ \boldsymbol{\mu}_{\phi^{(v)}}(\mathbf{x}^{(v)}),\ \boldsymbol{\Sigma}_{\phi^{(v)}}(\mathbf{x}^{(v)})),\end{aligned} \tag{4}$$

where λ_v denotes the non-negative normalized coefficient for the v-th component, satisfying $\lambda_v \geq 0$ and $\sum_{v=1}^{m} \lambda_v = 1$. The mean and covariance of multivariate Gaussian distribution w.r.t $\mathbf{z}$ can be obtained from the encoder with parameters $\phi^{(v)}$. Consequently, the lower bound on the evidence $p(x^{(v)})$ can be rewritten as,

$$\begin{aligned}\mathcal{L}_{\text{ELBO}} = & - D_{\text{KL}}\left(\sum_{v=1}^{m} \lambda_v q_{\phi^{(v)}}(\mathbf{z}|\mathbf{x}^{(v)})\|p(\mathbf{z})\right) \\ & + \sum_{v=1}^{m} \lambda_v \mathbb{E}_{q_{\phi^{(v)}}(\mathbf{z}|\mathbf{x}^{(v)})}\left[\sum_{j=1}^{m} \log p_{\theta^{(v)}}(\mathbf{x}^{(j)}|\mathbf{z})\right].\end{aligned}$$

Note that the KL divergence term is difficult to compute analytically, we turn to approximate its upper bound.

Lemma 1. *For all non-negative measurable functions* $f_i : \mathbb{R} \to [0, \infty)$ *satisfying* $\int f_i(\mathbf{x})d\mathbf{x} = 1$, *defining a weighting function* $g(\mathbf{x}) = \sum_i^m \lambda_i f_i(\mathbf{x})$, *with* $\lambda_i \geq 0$ *and* $\sum_i^m \lambda_i = 1$, one has

$$\int \sum_{i}^{m} \lambda_i f_i(\mathbf{x}) \log g(\mathbf{x})d\mathbf{x} \leq \sum_{i}^{m} \lambda_i \int f_i(\mathbf{x}) \log f_i(\mathbf{x})d\mathbf{x}.$$

Proof. Let f_i be a probability density function, and $g(\mathbf{x})$ be a mixture of density with m components of f_i. Consider $\mathbb{E}_{f_i(\mathbf{x})}\left[\log \frac{g(\mathbf{x})}{f_i(\mathbf{x})}\right]$, by utilizing Jensen's inequality, we have,

$$\mathbb{E}_{f_i(\mathbf{x})}\left[log\frac{g(\mathbf{x})}{f_i(\mathbf{x})}\right] \leq \mathbb{E}\left[f_i(\mathbf{x})\frac{g(\mathbf{x})}{f_i(\mathbf{x})}\right] = \log\left(\int g(\mathbf{x})d\mathbf{x}\right) = 0.$$

Thus, it can be showed that $\int f_i(\mathbf{x}) \log g(\mathbf{x})d\mathbf{x} \leq \int f_i(\mathbf{x}) \log f_i(\mathbf{x})d\mathbf{x}$.

With Lemma 1, the new objective of VMVAE is formulated by,

$$\begin{aligned}\mathcal{L}_{\text{VMVAE}} = & - \sum_{v=1}^{m} \lambda_v D_{\text{KL}}\left(q_{\phi^{(v)}}(\mathbf{z}|\mathbf{x}^{(v)})\|p(\mathbf{z})\right) \\ & + \sum_{v=1}^{m} \lambda_v \mathbb{E}_{q_{\phi^{(v)}}(\mathbf{z}|\mathbf{x}^{(v)})}\left[\sum_{j=1}^{m} \log p_{\theta^{(v)}}(\mathbf{x}^{(j)}|\mathbf{z})\right] \\ & \leq \mathcal{L}_{\text{ELBO}}.\end{aligned} \tag{5}$$

The proposed VMVAE is enlightened by Eq. (5). The summation of KL-divergences drives the unimodal variational posteriors to approach the prior individually, while the summation of their expectations reveals alternatively variational inference followed by full view generation. Notably, the greater expectation of v-th view over all views, indicating the more complete information it contains, and therefore the greater the weight λ_v. The whole model is illustrated in Fig. 1(a).

3.2 Complete Multi-view VAE

The proposed vanilla multi-view VAE faces a pitfall when dealing with missing views. When there are m views, 2^m inference paths need to be constructed [15] and thus render the method intractable as the number of views increases. Either the inference about missing views is abandoned, which would be biased by the view weights and cause a large ELBO slack. To tackle this issues, one can exploit the intrinsic correlations between views, thereby using the extracted representation to engineer a learning function to deal with incomplete multi-view problem. Let $(\{\mathbf{z}^{(v)}\}, \mathbf{c})$ denote the view-peculiar and complete generative latent variables. The two type of latent variable can be modeled by a linear transformation $\mathbf{z}^{(w)} = \mathbf{z}^{(v)}C_{vw},\ C_{vw} \in \mathbb{R}^{d_{\mathbf{z}} \times d_{\mathbf{z}}}$. This correlation enables interconversion in linear spaces of the same dimension, and will not change due to unobservable views. For a random variable obeying the Gaussian distribution, given $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$, whose linear transformation distribution is $\mathbf{y}C \sim \mathcal{N}(\boldsymbol{\mu}C,\ C^T\boldsymbol{\Sigma}C)$ under the statistical principle. It is noted that, similar to work [41], under the assumption of nonlinearity, correlations between views can be more flexible and fitted by using neural networks.

In this way, we introduce a novel joint variational posterior as,

$$\begin{aligned} & q(\{\mathbf{z}^{(v)}\}, \mathbf{c}|\{\mathbf{x}^{(v)}\}) \\ &= \sum_{v=1}^{m} \lambda_v q(\{\mathbf{z}^{(v)}\}, \mathbf{c}|\mathbf{x}^{(v)}) \\ &= \sum_{v=1}^{m} \lambda_v q_\psi(\mathbf{c}|\{\mathbf{z}^{(v)}\}) \prod_{w \neq v}^{m} q(\mathbf{z}^{(w)}|\mathbf{z}^{(v)}) q_{\phi^{(v)}}(\mathbf{z}^{(v)}|\mathbf{x}^{(v)}), \end{aligned} \tag{6}$$

where $q(\mathbf{z}^{(w)}|\mathbf{z}^{(v)})q_{\phi^{(v)}}(\mathbf{z}^{(v)}|\mathbf{x}^{(v)}) = \mathcal{N}(\mathbf{z}^{(w)};\ \boldsymbol{\mu}_{\phi^{(v)}}(\mathbf{x}^{(v)})C_{vw},\ C_{vw}^T\boldsymbol{\Sigma}_{\phi^{(v)}}(\mathbf{x}^{(v)})C_{vw})$. Here the generative latent variable $q_\psi(\mathbf{c}|\{\mathbf{z}^{(v)}\})$ is obtained from multiple view-peculiar variables under the fusion network with parameter ψ, i.e., $\mathbf{c} \sim \mathcal{N}(\boldsymbol{\mu}_\psi(\{\mathbf{z}^{(v)}\}),\ \boldsymbol{\Sigma}_\psi(\{\mathbf{z}^{(v)}\}))$.

Under the joint inference model (6), the framework of complete multi-view VAE is shown in Fig. 1(b). All view-peculiar encoders are optimized to learn a balanced representation, and the fusion network is leveraged to aggregate information from all views. When one view is missing, the latent representation can be predicted by other available view information and learned invariant correlations. For example, given an available view v and an unobservable view u, the inference process for view u can be formulated by,

$$\begin{aligned} & \lambda_u q(\{\mathbf{z}^{(v)}\}, \mathbf{c}|\mathbf{x}^{(u)}) \\ &= \lambda_u q_\psi(\mathbf{c}|\{\mathbf{z}^{(v)}\}) \prod_{w \neq u}^{m} q(\mathbf{z}^{(w)}|\mathbf{z}^{(u)}) q(\mathbf{z}^{(u)}|\mathbf{z}^{(v)}) q_{\phi^{(v)}}(\mathbf{z}^{(v)}|\mathbf{x}^{(v)}). \end{aligned} \tag{7}$$

Eq. (7) shows how the weights of the learned distributions remain valid once the latent variables of the unobservable views are modeled.

To correctly model the correlations between view-peculiar latent variable, it is desired to approximate the true transformations by minimizing the KL-divergence,

$$\begin{aligned} & \min D_{\mathrm{KL}}\left(p(\mathbf{z}^{(w)}|\mathbf{z}^{(v)})\|q(\mathbf{z}^{(w)}|\mathbf{z}^{(v)})\right) \\ &= \max \mathbb{E}_{p(\mathbf{z}^{(w)}|\mathbf{z}^{(v)})}\left[\log q(\mathbf{z}^{(w)}|\mathbf{z}^{(v)})\right]. \end{aligned} \tag{8}$$

Therefore, the objective of CMVAE can be rewritten as,

$$\begin{aligned} \mathcal{L}_{\text{CMVAE}} = & -\sum_{v=1}^{m} \lambda_v D_{\mathrm{KL}}\left(q(\{\mathbf{z}^{(v)}\}, \mathbf{c}|\mathbf{x}^{(v)})\|p(\mathbf{c})\right) \\ & + \sum_{v=1}^{m} \lambda_v \mathbb{E}_{q(\{\mathbf{z}^{(v)}\}, \mathbf{c}|\mathbf{x}^{(v)})}\left[\sum_{j=1}^{m} \log p_{\theta^{(v)}}(\mathbf{x}^{(j)}|\mathbf{c})\right] \\ & + \sum_{v=1}^{m}\sum_{w \neq v}^{m} \mathbb{E}_{p(\mathbf{z}^{(w)}|\mathbf{z}^{(v)})}\left[\log q(\mathbf{z}^{(w)}|\mathbf{z}^{(v)})\right]. \end{aligned} \tag{9}$$

Note that the third term of Eq. (9) calculates only the available paired views.

Lemma 2. *For any multivariate random variable $\mathbf{x} \in \mathbb{R}^J$, and its density function $p(\mathbf{x})$, given $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \mathbf{I})$, we have,*

$$\mathbb{E}_{p(\mathbf{x})}\left[\log q(\mathbf{x})\right] \leq -\frac{J}{2}\log 2\pi < 0.$$

Proof. By using Monte Carlo estimation to approximate the expectation, we take T samples of $\mathbf{x}_t$ from the density $p(\mathbf{x})$,

$$\begin{aligned} & \mathbb{E}_{p(\mathbf{x})}\left[\log q(\mathbf{x})\right] \\ &= \frac{1}{T}\sum_{t=1}^{T} \log \frac{1}{\sqrt{(2\pi)^J}} \exp(-\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu})^T(\mathbf{x}_t - \boldsymbol{\mu})) \\ &= -\frac{J}{2}\log 2\pi - \frac{1}{2T}\sum_{t=1}^{T} \|\mathbf{x}_t - \boldsymbol{\mu}\|_2^2 \\ &\leq -\frac{J}{2}\log 2\pi < 0. \end{aligned}$$

With Lemma 2, by simply setting the covariance matrix of $q(\mathbf{z}^{(w)}|\mathbf{z}^{(v)})$ as identity matrix, the objective of CMVAE can be seen as the lower bound of the joint likelihood function, i.e., $\log p(\{\mathbf{x}^{(v)}\}) \geq \mathcal{L}_{\text{ELBO}}(\{\mathbf{x}^{(v)}\}) \geq \mathcal{L}_{\text{CMVAE}}(\{\mathbf{x}^{(v)}\})$. Finally, the maximization of the joint likelihood function is converted into the maximization of the CMVAE objective.

3.3 Numerical Scheme to Solve CMVAE

The Eq. (9) characterize an unified objective function for optimizing the parameters of view-peculiar encoders, pairwise correlation matrices, fusion network and multiple decoders. For optimization using Stochastic Gradient Variational Bayes (SGVB), the sampling operations of $(\{\mathbf{z}^{(v)}\}, \mathbf{c})$ should be mapped to the deterministic functions. By the *reparameterization trick* [35] for the continuous variable, we sample the t-th latent representations by,

$$\mathbf{z}_t^{(v)} = \boldsymbol{\mu}_{\phi^{(v)}} + \boldsymbol{R}_{\phi^{(v)}}\boldsymbol{\epsilon}_t^{(v)} \tag{10}$$

$$\mathbf{c}_t = \boldsymbol{\mu}_\psi + \boldsymbol{R}_\psi\boldsymbol{\epsilon}_t \tag{11}$$

where $\boldsymbol{R}_{\phi^{(v)}}\boldsymbol{R}_{\phi^{(v)}}^T = \boldsymbol{\Sigma}_{\boldsymbol{\phi^{(v)}}},\ \boldsymbol{\epsilon}_t^{(v)} \in \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{R}_\psi\boldsymbol{R}_\psi^T = \boldsymbol{\Sigma}_\psi,\ \boldsymbol{\epsilon}_t \in \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Algorithm 1 Optimization Procedure of CMVAE

Input: Multi-view dataset $\boldsymbol{\mathcal{X}}$; Statistical model of the prior $p(\mathbf{c}) = \mathcal{N}(\mathbf{0}, \boldsymbol{I})$; Setting $T = 1$ and the dimensionality of latent variables.

Parameter: Initialize parameters $\{\phi^{(v)}\}$, $\{\theta^{(v)}\}$, ψ with random values, $\lambda_v = \frac{1}{m}$ and C_{vw} with identity matrix.

while not reaching the maximal epochs **do**
  for v in m views **do**
    Calculate $(\boldsymbol{\mu}_{\phi^{(v)}}(\mathbf{x}^{(v)}),\ \boldsymbol{\Sigma}_{\phi^{(v)}}(\mathbf{x}^{(v)}))$ through v-th encoder and then sample $\mathbf{z}_t^{(v)}$ by Eq. (10);
    Implement $\mathbf{z}^{(w)} = \mathbf{z}^{(v)} C_{vw}, \forall w \in 1, 2, ..., m,\ w \neq v$;
    Calculate $(\boldsymbol{\mu}_{\psi}(\{\mathbf{z}^{(v)}\}),\ \boldsymbol{\Sigma}_{\psi}(\{\mathbf{z}^{(v)}\}))$ through the fusion network and then sample $\mathbf{c}_t$ by Eq. (11);
  for j in m views **do**
    Generate $\{\mathbf{x}^{(v)}\}$ by m decoders.
  end for
  end for
  Update $\{\phi^{(v)}\}$, $\{\theta^{(v)}\}$, ψ, C_{vw}, λ_v by maximizing Eq. (12).
end while

Output: The complete generative latent representation $\mathbf{c}$.

By using Monte-Carlo estimators, the objective of CMVAE can be further written as,

$$\begin{aligned}
&\mathcal{L}_{\text{CMVAE}}(\{\mathbf{x}^{(v)}\}) \\
=&\sum_{v=1}^{m} \frac{\lambda_v}{T} \sum_{t=1}^{T} \Bigg[\sum_{j=1}^{m} \log p_{\theta^{(v)}}(\mathbf{x}^{(j)}|\mathbf{c}_t) - \log q_{\psi}(\mathbf{c}_t|\{\mathbf{z}_t^{(v)}\}) \\
&- \sum_{w \neq v}^{m} \log q(\mathbf{z}_t^{(w)}|\mathbf{z}_t^{(v)}) - \log q_{\phi^{(v)}}(\mathbf{z}_t^{(v)}|\mathbf{x}^{(v)}) + \log p(\mathbf{c}_t) \Bigg] \\
&- \frac{1}{2} \sum_{v=1}^{m} \sum_{w \neq v}^{m} \|\mathbf{z}_t^{(w)} - \boldsymbol{\mu}_{\phi^{(v)}}(\mathbf{x}^{(v)}) C_{vw}\|_2^2, \qquad (12)
\end{aligned}$$

where T denotes the number of Monte Carlo samples and is usually set to be 1. The partial derivatives of each parameter combination are then calculated and used in the stochastic back-propagation technique. The partial derivatives of each parameter combination are then calculated as follows,

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}^{(\boldsymbol{v})}} = \sum_{j=1}^{m} \frac{\lambda_j}{T} \sum_{t=1}^{T} \frac{\partial}{\partial \theta^{(v)}} \log p_{\theta^{(v)}}(\mathbf{x}^{(v)}|\mathbf{c}_t^{(j)}), \qquad (13)$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{\psi}} =& \sum_{v=1}^{m} \frac{\lambda_v}{T} \sum_{t=1}^{T} \frac{\partial}{\partial \mathbf{c}_t^{(v)}} \Bigg[\sum_{j=1}^{m} \log p_{\theta^{(v)}}(\mathbf{x}^{(j)}|\mathbf{c}_t^{(v)}) \\
&+ \log p(\mathbf{c}_t^{(v)}) \Bigg] \cdot \left(\frac{\partial \boldsymbol{\mu}_{\psi}}{\partial \psi} + \frac{\partial \boldsymbol{R}_{\psi}}{\partial \psi} \cdot \boldsymbol{\epsilon}_t \right) \\
&- \frac{\partial}{\partial \psi} \log q_{\psi}(\mathbf{c}_t^{(v)}|\{\mathbf{z}_t^{(v)}\}), \qquad (14)
\end{aligned}$$

TABLE 1
Statistics on tested nine datasets

Datasets	# Samples	# Views	# Classes	Dimensionality
MSRC-V1	240	5	7	24,576,512,256,254
Notting-Hill	550	3	5	2000,3304,6750
Handwritten	2000	5	10	240,76,216,47,64
Caltech101-20	2386	6	20	48,40,254,1984,512,928
BDGP	2500	2	5	1750,79
Animal	10158	2	50	4096,4096
PolyMNIST	60000	5	10	784,784,784,784,784
Multiome PBMC	11909	2	11	36601,108377
Multiome BMNC	69249	2	22	13431,116490

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial C_{vw}} =& \frac{\lambda_v}{T} \sum_{t=1}^{T} \frac{\partial}{\partial \mathbf{c}_t^{(v)}} \Bigg[\sum_{j=1}^{m} \log p_{\theta^{(v)}}(\mathbf{x}^{(j)}|\mathbf{c}_t^{(v)}) + \log p(\mathbf{c}_t^{(v)}) \Bigg] \\
&\left(\frac{\partial \boldsymbol{\mu}_{\psi}}{\partial \mathbf{z}_t^{(w)}} + \frac{\partial \boldsymbol{R}_{\psi}}{\partial \mathbf{z}_t^{(w)}} \cdot \boldsymbol{\epsilon}_t \right) \cdot \mathbf{z}_t^{(v)} - \frac{\partial}{\partial C_{vw}} \log q(\mathbf{z}_t^{(w)}|\mathbf{z}_t^{(v)}) \\
&+ (\mathbf{z}_t^{w} - \boldsymbol{\mu}_{\phi^{(v)}}(\mathbf{x}^{(v)}) C_{vw}) \cdot \boldsymbol{\mu}_{\phi^{(v)}}(\mathbf{x}^{(v)}), \qquad (15)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{\phi}^{(\boldsymbol{v})}} =& \frac{\lambda_v}{T} \sum_{t=1}^{T} \frac{\partial}{\partial \mathbf{c}_t^{(v)}} \Bigg[\sum_{j=1}^{m} \log p_{\theta^{(v)}}(\mathbf{x}^{(j)}|\mathbf{c}_t^{(v)}) + \log p(\mathbf{c}_t^{(v)}) \Bigg] \\
&\left(\frac{\partial \boldsymbol{\mu}_{\psi}}{\partial \mathbf{z}_t^{(v)}} + \frac{\partial \boldsymbol{R}_{\psi}}{\partial \mathbf{z}_t^{(v)}} \cdot \boldsymbol{\epsilon}_t \right) \cdot \left(\frac{\partial \boldsymbol{\mu}_{\phi^{(v)}}}{\partial \phi^{(v)}} + \frac{\partial \boldsymbol{R}_{\phi^{(v)}}}{\partial \phi^{(v)}} \cdot \boldsymbol{\epsilon}_t^{(v)} \right) \\
&- \frac{\partial}{\partial \phi^{(v)}} \log q_{\phi^{(v)}}(\mathbf{z}_t^{(v)}|\mathbf{x}^{(v)}), \qquad (16)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \lambda_v} =& \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{m} \log p_{\theta^{(j)}}(\mathbf{x}^{j}|\mathbf{c}_t^{(v)}) \\
&- \log q(\{\mathbf{z}_t^{(v)}\}, \mathbf{c}_t^{(v)}|\mathbf{x}^{(v)}) + \log p(\mathbf{c}_t^{(v)}). \qquad (17)
\end{aligned}$$

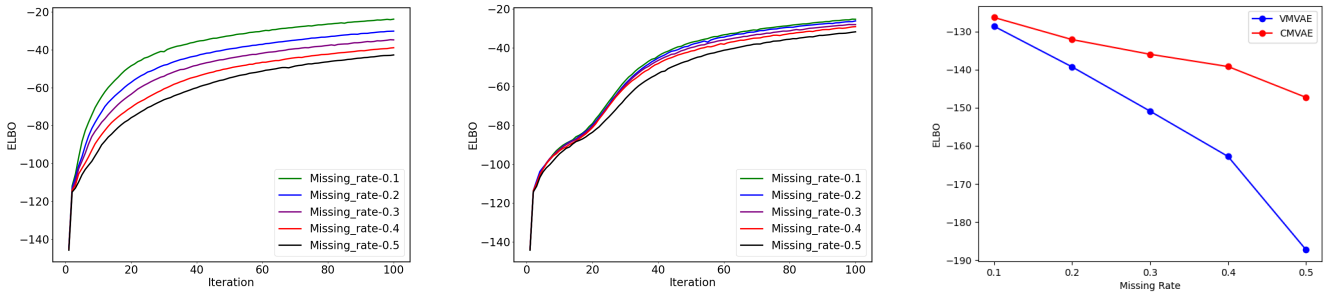
The concrete optimization procedure of CMVAE is summarized in Algorithm 1.

4 Experimental Results

We evaluate the effectiveness of multi-view latent representations based on three different aspects. The approximation of the joint data distribution is measured in terms of log-likelihood. The **completeness** is assessed by the clustering task and classification task compared to state-of-the-art incomplete multi-view learning algorithms. The **cross-generative** of the latent representation is demonstrated by qualitative results of cross-view image generation. In addition, the efficiency and stability of the proposed model are verified by time complexity and parameter sensitivity analysis. Finally, the practical significance of the model is demonstrated by applying it on real-world generated bioinformatic data.

4.1 Experimental Settings

Model setup: For the clustering and classification task, the architectures of $q_{\phi^{(v)}}(\mathbf{z}^{(v)}|\mathbf{x}^{(v)})$ and $p_{\theta^{(v)}}(\mathbf{x}^{(v)}|\mathbf{c})$ are fully



(a) VMVAE-Handwritten (b) CMVAE-Handwritten (c) Caltech101-20

Fig. 2. The variation of the objective values in terms of training iteration for (a) VMVAE, and (b) CMVAE, on the Handwritten dataset. The convergence values reached by ELBO decrease as the missing rate increases, while CMVAE ultimately achieves a higher ELBO value. (c) The pronounced difference in ELBO values on the Caltech101-20 dataset verifies that CMVAE has a tighter lower bound, especially with large amounts of information missing.

connected networks with d_v-500-500-1204-256 and 256-1024-500-500-d_v neurons, respectively, where d_v is the dimensionality of each view. The fusion network $q_{(\psi)}(\mathbf{c}|\{\mathbf{z}^{(v)}\})$ concatenates multiple view-peculiar latent representations, followed by a fully connected layer with dimensionality D. For cross-view image generation, we set up the encoders and decoders as CNN network, whose specific architecture follows MoPoE-VAE [15]. Adam optimizer [42] is utilized to maximize the objective function, and set the learning rate to be 0.001 with a decay of 0.9 for every 10 epochs.

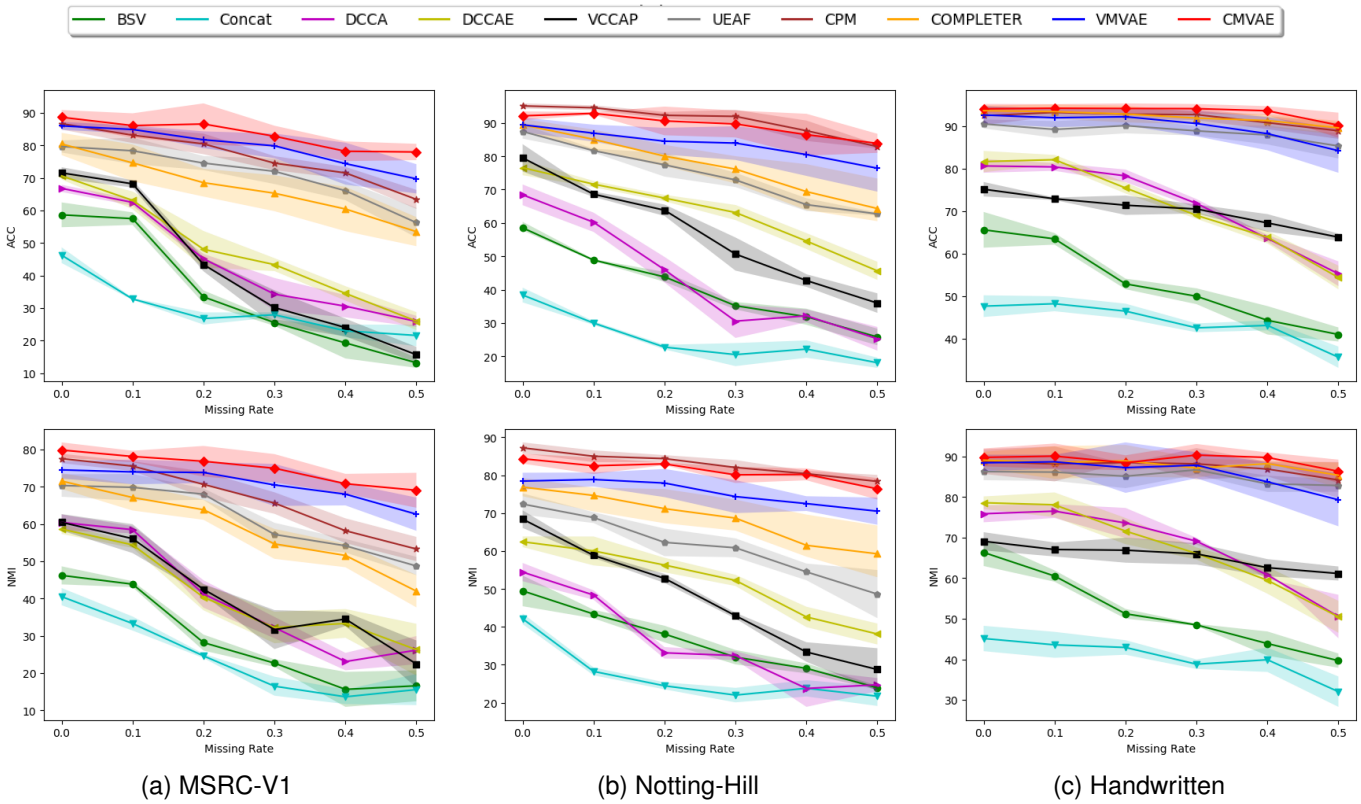
Datasets: We adopt the following seven benchmark datasets ranging from small samples to large-scale and two real-world bioinformatic datasets, and the detailed statistics are shown in Table 1.

- **MSRC-V1** [43] consists of 240 images and 9 object categories. We select 7 of the whole object classes, namely tree, building, airplane, cow, face, car and bicycle, and extract HOG, LBP features as 2 views.
- **Notting-Hill** [44] is widely used video face dataset for clustering, which collects 4660 faces across 76 tracks of the 5 main actors from the movie 'Notting Hill'. We use the multi-view version provided in [45], consisting of 550 images with three kind of features, i.e., LBP, gray pixels, and Gabor features.
- **Handwritten digit** [46] contains 2000 samples with 10 numerals from 0 to 9 with five views which are respectively extracted by Fourier coefficients, profile correlations, Karhunen-Love coefficient, Zernike moments, and pixel average extractors.
- **Caltech101-20** is a subset of the object recognition dataset [47] containing 20 classes and 6 different views with a total of 2386 samples, including Gabor features, wavelet moments, CENTRIST features, histogram of oriented gradients, GIST features and local binary patterns.
- **BDGP** [48] contains 2500 images in 5 categories, and each sample is described by a 1750-D image vector and a 79-D textual feature vector.
- **Animal** [49] contains 10158 animal images divided into 50 categories. Two types of deep features extracted by [50] and [51] respectively are considered as two views.
- **PolyMNIST** [15] dataset consists of a total of 60,000 samples of 5 different MNIST images that have different backgrounds and writing styles, but have the same numerical labels.

Compared algorithms: Two baselines and six state-of-the-art algorithms are used to compare the clustering performance, including:

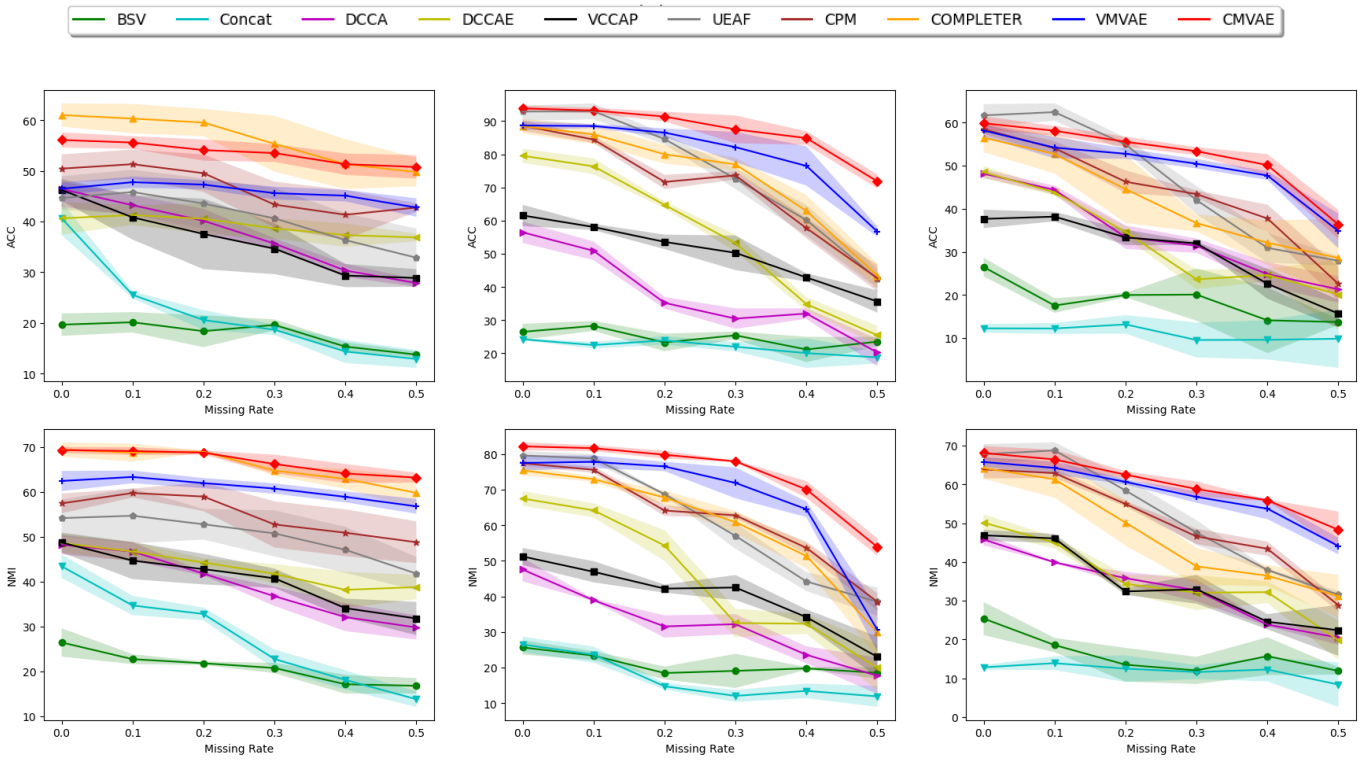
- **Best Single View (BSV)** selects the best k-means clustering results among all single views.
- **Concat** method stacks the features of all views and conducting k-means clustering on it.
- **DCCA** [2] extracts flexible nonlinear representations with respect to the correlation objective measured on two views data.
- **DCCAE** [22] extends DCCA by using autoencoders to extract common low-dimensional embeddings, and jointly optimizes the correlation objective and reconstruction loss.
- **VCCAP** [3] leverages deep generative models to implement a natural idea that multiple views can be generated from a shared latent variable.
- **UEAF** [52] reconstructs the hidden information of missing views with preserving the local structure, and considers the adaptive importance of different views.
- **CPM** [17] learns a unified latent representation by jointly considering completeness and structure, which is highly flexible and generalizable to incomplete multi-view data.
- **COMPLETER** [33] learns informative and consistent representations by maximizing the mutual information between different views, and recovers missing views by minimizing the conditional entropy of different views through dual prediction.

Incomplete data construction: To preprocess the dataset according to the settings in [17], we set the missing rate $\eta = \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$, then randomly selected $\eta \times n \times m$ samples as missing data. Then random instances were removed from each view, in the case that all samples were guaranteed to retain at least one view.



(a) MSRC-V1 (b) Notting-Hill (c) Handwritten

Fig. 3. Clustering performance comparison in terms of NMI and Accuracy by tested ten methods under different missing rates, on (a) MSRC-V1, (b) Notting-Hill, and (c) Handwritten.



(a) Caltech101-20 (b) BDGP (c) Animal

Fig. 4. Clustering performance comparison in terms of NMI and Accuracy by tested ten methods under different missing rates, on (a) Caltech101-20, (b) BDGP, and (c) Animal.

TABLE 2
Classification accuracy comparison under different missing rate on three datasets (mean±standard deviation). Higher values indicate better performance. The optimal and suboptimal results are in bold and underlined, respectively.

Datasets	Methods	0	0.1	0.2	0.3	0.4	0.5
MSRC-V1	BSV	72.84±1.21	70.95±1.62	63.30±3.27	55.70±3.15	49.28±3.41	40.01±4.55
	Concat	69.47±2.84	55.54±3.06	45.43±5.04	40.31±5.41	38.34±4.60	34.79±6.32
	DCCA [2]	74.23±2.65	72.00±4.54	62.16±4.55	56.97±6.64	41.67±8.86	30.78±3.43
	DCCAE [22]	80.46±3.02	78.20±3.46	70.82±4.62	65.47±5.89	59.36±6.14	50.12±5.18
	VCCAP [3]	81.24±2.84	70.56±2.54	62.80±3.58	57.20±5.57	50.77±7.21	44.49±5.07
	UEAF [52]	85.45±1.54	85.67±3.07	83.63±4.05	80.86±6.19	73.96±4.62	70.83±6.20
	CPM [17]	90.14±1.08	88.69±1.28	87.14±2.98	**87.50±2.47**	77.90±4.81	72.38±5.08
	COMPLETER [33]	88.16±2.78	83.65±3.55	80.78±4.84	78.15±4.74	76.34±6.47	69.20±5.65
	VMVAE	89.45±1.45	88.20±3.60	85.71±3.56	80.20±4.85	77.07±3.97	72.91±5.92
	CMVAE	**90.86±0.94**	**90.47±2.26**	**88.09±2.34**	87.45±3.91	**85.45±4.37**	**81.85±4.48**
Notting-Hill	BSV	64.45±1.87	66.22±1.21	61.65±1.37	51.40±2.36	44.17±2.79	37.79±3.00
	Concat	48.14±3.10	35.97±0.96	30.95±1.52	28.65±2.57	26.11±1.27	24.93±2.94
	DCCA [2]	71.54±1.17	69.50±0.77	58.91±1.61	50.46±3.92	41.48±3.29	37.30±5.15
	DCCAE [22]	78.45±1.24	76.48±1.18	74.47±1.77	71.28±2.57	64.47±1.63	62.72±2.39
	VCCAP [3]	77.59±2.14	73.37±0.48	69.17±1.06	64.71±2.47	60.37±1.68	58.47±2.05
	UEAF [52]	89.54±1.78	84.67±1.95	82.31±1.74	81.84±1.36	78.77±2.77	73.57±2.35
	CPM [17]	**97.28±1.65**	**97.77±1.10**	**97.66±1.31**	**96.72±2.12**	**96.40±3.10**	**96.06±2.20**
	COMPLETER [33]	96.86±1.45	95.15±1.32	95.04±1.42	92.57±3.45	93.08±2.72	92.17±2.75
	VMVAE	90.08±2.87	89.55±3.35	88.52±2.18	85.81±1.73	84.35±1.88	81.02±1.91
	CMVAE	96.12±0.81	94.81±1.71	94.61±1.69	94.37±1.14	93.37±1.41	93.19±1.73
Handwritten	BSV	83.15±0.35	82.59±0.45	76.82±0.65	65.82±0.58	59.31±1.55	51.40±1.42
	Concat	94.65±1.04	95.32±0.53	93.82±0.86	92.60±0.66	90.25±0.69	87.47±1.12
	DCCA [2]	85.45±1.36	78.58±2.54	68.90±1.65	60.25±2.19	51.44±1.45	37.65±2.15
	DCCAE [22]	91.27±1.81	88.26±1.56	80.26±1.36	70.45±2.48	59.34±2.45	50.06±1.89
	VCCAP [3]	71.54±2.74	64.83±1.44	60.15±1.85	50.83±4.57	45.41±3.60	38.59±4.88
	UEAF [52]	93.64±1.98	92.48±1.47	92.85±1.83	92.23±2.07	92.22±2.46	91.57±1.78
	CPM [17]	94.78±0.48	94.82±0.85	93.67±1.70	93.56±1.84	92.66±2.40	91.03±2.10
	COMPLETER [33]	96.18±1.48	95.68±1.30	95.45±1.02	93.81±2.05	92.51±1.67	91.91±2.54
	VMVAE	95.23±1.68	94.45±1.21	94.21±2.20	93.53±2.60	93.18±2.30	92.64±3.05
	CMVAE	**96.69±0.65**	**96.25±0.35**	**96.24±1.31**	**95.75±1.48**	**95.50±1.45**	**94.82±1.65**

4.2 Joint Likelihood Approximation

The value of the variational lower bounds affects the portrayal of the data distribution, as well as the accuracy of the inference of the posterior. This conclusion can be derived from the following equation,

$$\begin{aligned}
&\log p(\{\mathbf{x}^{(v)}\}) \\
&= D_{\mathrm{KL}}(q(\mathbf{z}|\{\mathbf{x}^{(v)}\})\|p(\mathbf{z}|\{\mathbf{x}^{(v)}\})) + \mathcal{L}_{\mathrm{VMVAE}} \\
&= D_{\mathrm{KL}}(q(\{\mathbf{z}^{(v)}\},\mathbf{c}|\{\mathbf{x}^{(v)}\})\|p(\{\mathbf{z}^{(v)}\},\mathbf{c}|\{\mathbf{x}^{(v)}\})) \\
&\quad - \sum_{v=1}^{m}\sum_{w\neq v}^{m}\mathbb{E}_{p(\mathbf{z}^{(w)}|\mathbf{z}^{(v)})}\left[\log q(\mathbf{z}^{(w)}|\mathbf{z}^{(v)})\right] + \mathcal{L}_{\mathrm{CMVAE}}.
\end{aligned}$$

It can be seen that when the variational lower bound is larger, the smaller the KL divergence term is, which means that the variational posterior is closer to the true posterior. We conduct experiments on the datasets Handwritten and Caltech101-20 for VMVAE and CMVAE, respectively, and the results are shown in Fig. 2. By observing the results, three conclusions can be drawn: i) The ELBO of both VMVAE and CMVAE decrease to different degrees as the missing rate increases, which verifies that the more difficult it is to estimate the joint data distribution as the missing rate increases. ii) CMVAE converges more slowly in the initial stage, which is caused by the increased complexity of the posterior inference process, but the ELBO value is larger than that of VMVAE in the final stage, which indicates that the posterior inference of CMVAE is better than that of VMVAE. iii) CMVAE is less sensitive to the missing rate, which can be seen more clearly in Fig. 2(c). This demonstrates that learning view-invariant information has a facilitating solidative on latent representation learning in the case of missing views.

4.3 Clustering Performance Evaluation

To further verify the effectiveness of the learning of latent representation by VMVAE and CMVAE, we conduct k-means directly on the latent representation $\mathbf{z}$ and $\mathbf{c}$, respectively.

TABLE 3
Classification accuracy comparison under different missing rate on three datasets (mean±standard deviation). Higher values indicate better performance. The optimal and suboptimal results are in bold and underlined, respectively.

Datasets	Methods	0	0.1	0.2	0.3	0.4	0.5
Caltech101-20	BSV	52.45±2.21	50.95±1.62	53.30±1.27	45.70±1.15	39.28±2.41	30.01±2.55
	Concat	67.91±2.08	65.54±1.06	55.43±1.04	50.31±1.41	45.34±0.60	32.79±0.32
	DCCA [2]	56.78±1.92	52.00±0.54	52.16±1.55	46.97±1.64	41.67±1.86	36.78±1.43
	DCCAE [22]	56.60±1.35	57.20±1.46	56.82±0.62	55.47±0.89	49.36±1.14	50.12±1.18
	VCCAP [3]	58.17±3.18	50.56±0.54	47.80±0.58	44.20±0.57	46.77±1.21	44.49±1.07
	UEAF [52]	76.15± 0.95	74.67±1.07	72.63±1.05	71.86±1.19	68.96±1.62	66.83±2.20
	CPM [17]	90.84±0.52	91.10±1.28	90.85±0.98	89.40±1.24	87.23±1.18	84.39±2.38
	COMPLETER [33]	91.48±0.84	89.65±0.55	88.68±0.84	86.15±1.74	85.14±1.47	84.80±1.65
	VMVAE	**92.58±0.78**	90.10±1.60	89.65±0.65	87.88±1.85	86.68±2.07	84.51±3.32
	CMVAE	92.48±0.65	**92.21±0.45**	**91.45±0.64**	**90.45±0.55**	**89.55±0.90**	**87.65±1.58**
BDGP	BSV	69.48±0.87	68.22±1.21	61.65±1.37	51.40±1.36	44.17±1.79	37.79±1.01
	Concat	68.45±1.77	62.97±0.96	50.95±1.52	41.65±1.57	36.11±2.27	34.93±2.94
	DCCA [2]	82.60±2.10	78.50±0.77	65.91±1.61	58.46±1.92	46.48±1.29	42.30±1.15
	DCCAE [22]	86.16±1.20	82.48±1.18	79.87±0.77	75.28±1.57	72.17±1.93	69.72±2.39
	VCCAP [3]	86.75±2.46	86.37±1.48	79.17±1.76	77.71±3.47	69.37±2.68	58.47±2.05
	UEAF [52]	96.42±0.58	93.67±1.25	90.11±1.74	87.84±1.36	85.77±2.77	82.57±3.35
	CPM [17]	96.12±0.75	94.90±1.40	92.06±0.91	88.52±2.12	85.12±1.19	78.56±1.20
	COMPLETER [33]	95.15±0.85	94.15±1.32	92.04±1.42	91.37±1.45	90.08±1.72	87.17±2.25
	VMVAE	98.28±0.32	97.45±0.43	96.60±0.68	95.40±0.80	94.22±1.20	91.20±1.80
	CMVAE	**98.58±0.45**	**98.60±0.40**	**98.12±0.45**	**97.60±0.60**	**95.80±1.10**	**93.20±1.60**
Animal	BSV	53.14±1.54	37.59±0.45	33.90±0.69	28.22±0.52	24.31±0.51	8.40±0.62
	Concat	76.78±0.85	74.52±0.83	70.82±0.89	67.40±1.26	60.52±1.66	57.47±1.15
	DCCA [2]	45.86±1.25	7.68±0.36	6.61±0.46	6.22±0.19	5.21±0.41	5.56±0.35
	DCCAE [22]	54.24±0.82	27.26±1.26	22.21±1.36	20.50±0.98	15.30±0.43	12.00±1.39
	VCCAP [3]	72.54±1.36	70.83±0.84	62.05±0.85	52.38±1.17	46.01±0.68	38.95±0.68
	UEAF [52]	85.65±1.10	82.04±1.49	77.35±1.48	74.68±2.07	72.49±2.44	69.70±2.74
	CPM [17]	85.14±0.58	83.52±1.55	77.98±1.70	74.86±1.90	73.06±1.49	70.73±2.16
	COMPLETER [33]	**86.45±0.64**	**84.16±1.30**	**80.27±2.02**	76.81±2.55	72.71±2.47	70.91±2.50
	VMVAE	82.54±1.35	79.78±0.81	77.21±0.76	75.45±1.23	72.94±1.81	69.54±2.15
	CMVAE	84.28±0.60	82.24±0.45	79.56±0.38	**77.65±0.78**	**74.65±1.05**	**72.12±1.16**

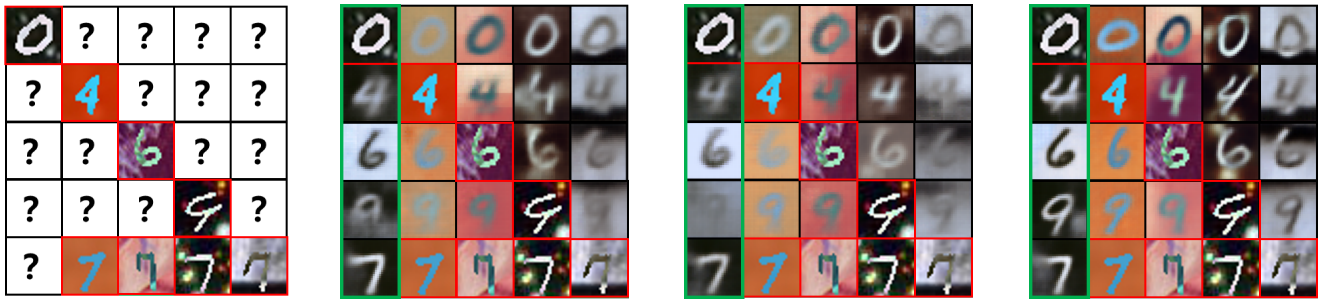
For BSV, Concat, DCCA, DCCAE, and VCCAP methods, we simply impute missing data as the mean of all samples in each view. Since CCA-based models can only handle two view data, we tested all two view combinations and finally reported the best clustering score.

For fairness, the parameter settings for the compared methods are done according to their authors' suggestions for their best clustering performances. All algorithms were replicated 10 times on the six datasets and the mean and standard deviation were recorded.

Evaluation metrics: For a comprehensive analysis, we use two popular clustering metrics including Normalized Mutual Information (NMI), Accuracy (ACC). The higher the values of these indicators, the better the clustering performance.

Clustering results and analysis: We tested ten methods on six multi-view datasets, suffering from different missing rates. The experimental results are summarized in Fig. 3 and Fig. 4. It can be observed that i) the multi-view learning approaches uniformly outperform the BSV and Concat clustering methods, especially when the samples are corrupted by missing views. The reason is that neither the BSV nor the Concat method exploits the relationship between different views. ii) CCA-based methods generally underperform than incomplete multi-view learning approaches, because filling missing values directly with the mean is inefficient, while targeted handling of missing data can more accurately mine missing view information. iii) The performance drop of all models is more pronounced as the missing rate increases for datasets with two views compared to datasets with three or more views. Because under the same missing rate, for samples with missing views, the fewer views have more missing information, making it difficult to recover complete view information.

By comparing the proposed model CMVAE with other incomplete multi-view clustering approaches, CMVAE is not the best performer when the view missing rate is small, but as the missing rate increases, the robustness of CMVAE is the best across six datasets. Taking the results of Caltech101-20 as an example, when $\eta = 0.1$, the ACC of



(a) Samples with missing view (b) MoPoE-VAE (c) VMVAE (d) CMVAE

Fig. 5. Visualization on cross-view image generation. (a) For each sample of 0, 4, 6, and 9, there are only one view are observed, while the others are missing. For the sample 7, there are only one missing view. The observed samples are used for generating the remaining view images by (b) MoPoE-VAE, (c) VMVAE, and (d) CMVAE. As can be seen, CMVAE shows the best detail in terms of figures structure and background, which is clearly contrasted in the first view, highlighted by the green box.

CMVAE is 55.61%, while COMPLETER is the best score of 60.34%, and when $\eta = 0.5$, CMVAE achieves the best score of 50.78%, while COMPLETER drops to 49.78%. Besides, it can be observed that CMVAE has less fluctuation in clustering performance at each missing rate compared to other incomplete multi-view learning methods. This illustrates the improvement from a more complete representation is more pronounced compared to other learning techniques.

On the other hand, comparing VMVAE and CMVAE, it can be seen that the clustering performance of CMVAE degrades to a weaker extent than VMVAE as the missing rate increases. This side-by-side confirms the point made in Section 4.2 that CMVAE has a better posterior inference capability and extracting view invariant information plays a role.

4.4 Classification Performance Evaluation

In this section, we evaluate the effectiveness of VMVAE and CMVAE for classification task on six datasets with different missing rates. The multi-view unified latent representations $\mathbf{z}$ and $\mathbf{c}$ are respectively fed into fully connected layers with the softmax activator. Network parameters are jointly optimized by adding cross-entropy loss.

For conventional multi-view learning methods, missing views are filled with mean values based on available samples in the same class. The CCA-based model reported the best classification scores for two views.

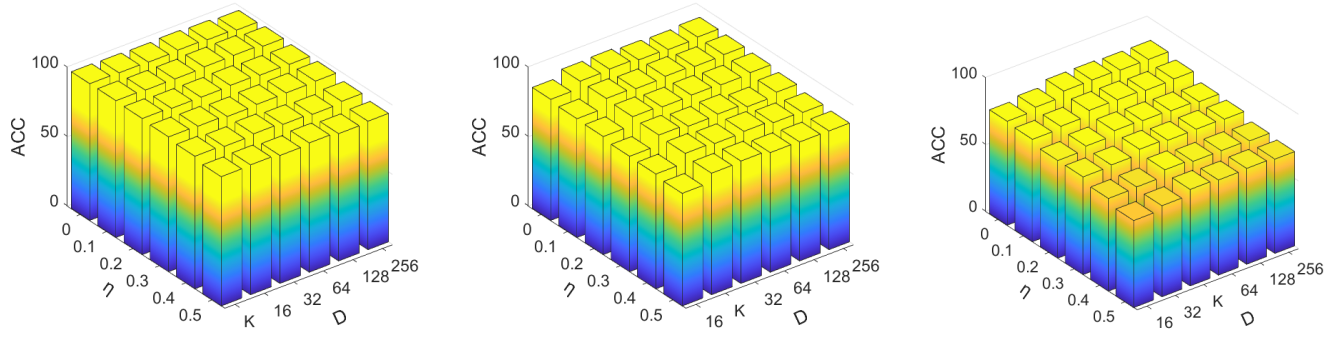
We divide 80% of the dataset as training set and 20% as testing set. All algorithms were repeated 10 times of five-fold cross-validation on six datasets according to the parameter settings suggested by the authors, and the mean and standard deviation of the accuracy were calculated.

Classification results and analysis: The experimental results are summarized in Table 2 and Table 3. Combined with the clustering results, the following three conclusions can be drawn: i) Compared with the conventional multi-view learning method, the incomplete multi-view learning method maintains advantages in the classification task under different missing rates, and due to the addition of labels, it is more robust to the missing view data than the clustering task. ii) CMVAE is still the best performing algorithm on the six datasets and the most robust to missing data. A total of 22 optimal performances and 5 sub-optimal performances were obtained across 30 classification metrics across six datasets. The second place is the CPM model, which has won 6 first places and 9 second places. It is worth noting that in the case of the Notting-Hill dataset and low missing rates, the adversarial strategy employed by CPM forces the generated data to obey the distribution of the observed data bringing an advantage to latent representation learning. On the other hand, when the missing rate is large, CMVAE has a large margin to lead. For example, when the missing rate is 0.5, CMVAE has an average accuracy advantage of 4.95% compared to CPM. iii) Compared with VMVAE, CMVAE also has a significant improvement in classification tasks. Combined with the results of clustering, it can be explained that mining the correlation between views and making full use of view invariant information is helpful for learning complete latent representations in the absence of views.

4.5 Cross-view Image Generation

To test the cross-generative of the latent representation, we conducted cross-view image generation experiments comparing MoPoE-VAE [15] on PolyMNIST dataset, and the qualitative results are shown in Fig. 5. We selected five digits $[0, 4, 6, 7, 9]$, which are more difficult to distinguish among handwritten digits, and constructed five training sets for each digit. MoPoE-VAE, VMVAE and CMVAE were trained on these five subsets to obtain five models for cross-view image generation under the following five given conditions: (1) digit 0 containing only view 1. (2) digit 4 containing only view 2. (3) digit 6 containing only view 3. (4) digit 9 containing only view 4 and (5) digit 7 containing views 2, 3, 4, 5.

It can be seen that in the first four cases, the quality of the images generated by CMVAE is significantly improved compared to MoPoE-VAE and VMVAE, as evidenced by the clarity of the data and the background details. The possible reason is that MoPoE-VAE poorly preserves view-specific factors of variation, while CMVAE captures the underline transformations of background, allowing learning a more complete latent representation. And comparing the first four cases to the fifth, there is a further improvement in the quality of the generation, suggesting that more complete view data can provide tighter evidence lower bounds. This



(a) BDGP (b) Caltech101-20 (c) Animal

Fig. 6. Classification accuracy of latent variable dimensions D under different missing rate η on (a) BGDP, (b) Caltech101-20, and (c) ANIMAL datasets. The number of classes is denoted by K.

TABLE 4
Time complexity and running time analysis for Handwritten dataset.

Methods	Time complexity	Running time /s
CPM [17]	$\mathcal{O}(nmdD + md)$	4.10±0.12
COMPLETER [33]	$\mathcal{O}(nmdD + m^2D^2)$	3.85±0.15
MoPoE-VAE [15]	$\mathcal{O}(n2^m dD + nmdD)$	28.45±1.20
VMVAE	$\mathcal{O}(nmdD)$	2.78±0.10
CMVAE	$\mathcal{O}(nmdD + m^2D^2)$	3.64±0.12

also confirms the benefit of CMVAE having a larger ELBO value than VMVAE.

4.6 Time Complexity and Parameter Sensitivity Analysis

The previous experiments on clustering, classification and cross-view image generation demonstrate the superiority and effectiveness of CMVAE in learning to generative yet complete multi-view representations from quantitative and qualitative perspectives, respectively. Furthermore, the efficiency and stability of CMVAE are illustrated by analyzing the time complexity and parameter sensitivity.

Time complexity. Denotes input data with batch size as n, the maximal dimension across all views as d and the dimension of the latent representation as D. The computational complexity of the encoder and decoder for m views is $\mathcal{O}(nmdD)$ for CPM, COMPLETER, VMVAE and CMVAE. The computational complexity of the discriminator for m views is $\mathcal{O}(md)$ for CPM, and the complexity of the latent variable transformations is $\mathcal{O}(m^2D^2)$ for CMVAE and COMPLETER. Due to the need to face meet view missing cases, MoPoE-VAE construct 2^m posteriori inferences with computational complexity $\mathcal{O}(n2^m dD)$. In general, $d > n \gg D > m$, so the theoretical time complexity ranking is MoPoE-VAE>CPM>COMPLETE=CMVAE>VMVAE.

Additionally, we tested the runtime for 20 iterations on the Handwritten dataset on a computer equipped with an NVIDIA® RTX 2070 GPU, where $n = 256$ and $D = 10$. The mean and standard deviation of the 10 tests are computed and summarized in Table 4.The results show that VMVAE has the fastest running time, followed by CMVAE, which is close to COMPLETER. The actual test runtime are generally consistent with the theoretical results.

Parameter sensitivity. To investigate the effect of the output neurons of fusion network $q_{(\psi)}(\mathbf{c}|\{\mathbf{z}^{(v)}\})$, i.e., the dimensionality of latent variable $\mathbf{c}$, on the classification accuracy. We chose three datasets BDGP, Caltech101-20 and Animal corresponding to cluster classes K of 5, 20, 50 respectively. The dimensionality of the latent variable were selected from $D = [K, 16, 32, 64, 128, 256]$ in turn, and classification test was performed at the missing rate of $\eta = [0, 0.1, 0.2, 0.3, 0.4]$, respectively.

The results are shown in Fig. 6, where it can be seen that no matter what the missing rate is, the classification accuracy changes only slightly when the dimensionality $D \geq K$, and decreases significantly when comparing $D < K$. This illustrates the stability of the model with respect to the dimensionality of the latent variables, and can also provide a basis for the operator to set the number of neurons, which is most directly done by setting $D = K$.

4.7 Application to Bioinformatic Data

The seven benchmark datasets analyzed in Table 1 rely on artificially incomplete multi-view data, where the proposed model achieves superior performance compared to state-of-the-art multi-view learning methods. This does not reflect the real-world situation, in which data may encounter measurement bias and flexible correlations between views. We therefore seek to demonstrate CMVAE on bioinformatic multi-omics data including Multiome PBMC and Multiome BMMC datasets. (1) **Multiome PBMC**[1]. Human peripheral blood mononuclear cell (PBMC) profiles generated by the 10× Genomics Multiome ATAC and RNA kit with 11,909 cells, measuring 36,601 genes and 108,377 open chromatin peaks simultaneously. (2) **Multiome BMMC** [53]. Single-cell multi-omics data collected from bone marrow mononuclear cells (BMMC) from 12 healthy human donors. Half of the samples were measured using paired RNA and ATAC kits,

1. https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k

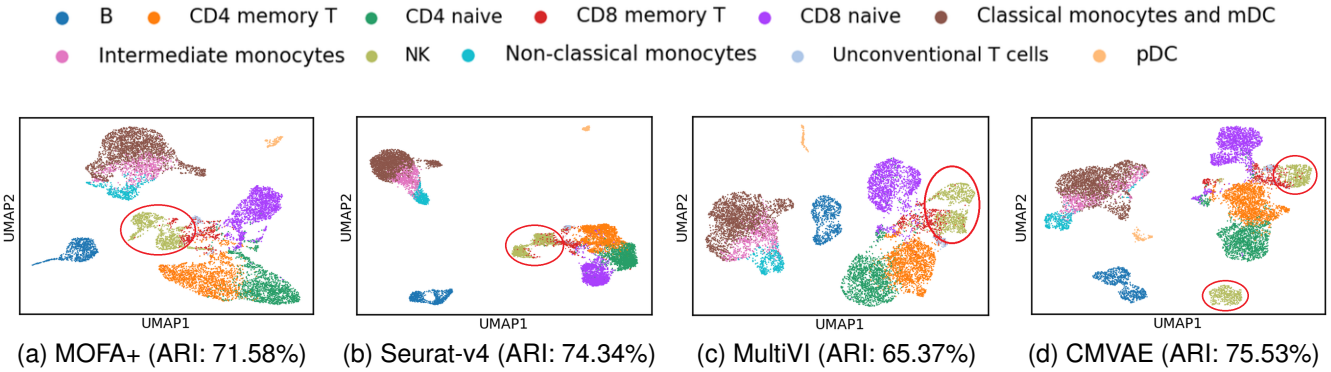


Fig. 7. Visualization results of multi-view latent representations using UMAP on Multiome PBMC dataset. Different colors represent different cell types. Through CMVAE, NK cells are more distinctly divided into two clusters in the embedding space, and the best cell typing performance for the ARI indicator is obtained.

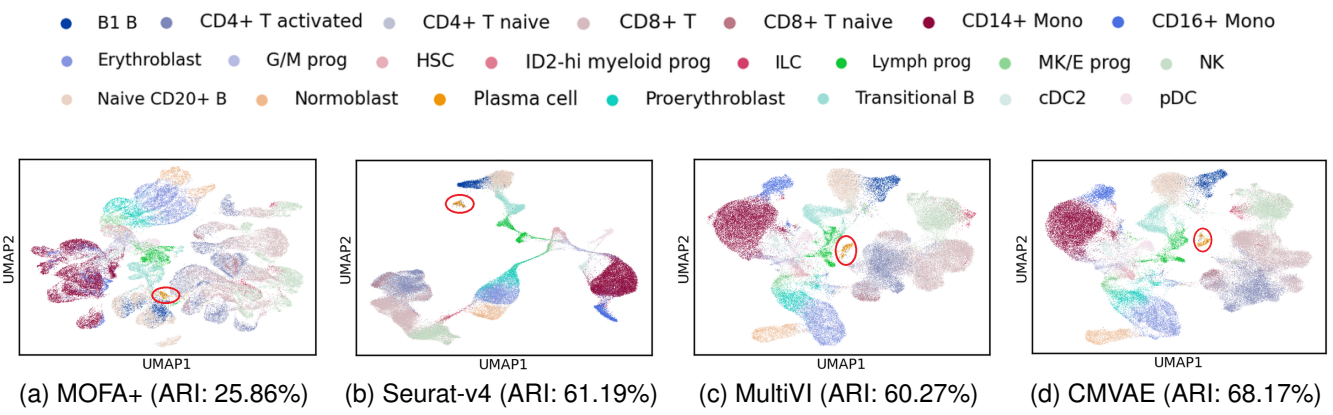


Fig. 8. Visualization results of multi-view latent representations using UMAP on Multiome BMMC dataset. Different colors represent different cell types. Through CMVAE, Plasma cell cluster can be more clearly separated, and the best cell typing performance for the ARI indicator is obtained.

and half were measured using single-cell gene expression kits only, for a total of 69,249 cells, 13,431 genes, and 116,490 open chromatin peaks. Quality control and preprocessing is performed on both datasets. Gene expression is filtered for high variant genes as 4,000 genes per cell and $\log(x+1)$ transformed. In addition, open chromatin peaks are binarized and 40,000 variable peaks are selected and log-normalized.

We compare CMVAE to other state-of-the-art cellular typing methods (1) **MOFA+** [54] utilizes computationally efficient variational inference to reconstruct low-dimensional representations of data and model variation in multi-omics single-cell genomic data. (2) **Seurat-v4** [55] introduces "weighted nearest neighbor" analysis to understand the relative utility of each genomic feature in each cell for comprehensive analysis of multi-omics data. (3) **MultiVI** [56] leverages three variational autoencoders for gene expression, chromatin accessibility, and protein abundance and estimates integrated cellular states by aligning and merging three modal latent states, driving missing view imputation from consistent information. For MOFA+ and MultiVI, we run with default parameters. In Seurat-v4, we first compute the Weighted Nearest Neighbor (WNN) graph, and then to obtain embeddings in the latent space, we run supervised PCA using the default parameters.

We directly perform Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [57] on the latent variables learned by MOFA+, Seurat-v4, MultiVI and CMVAE. Fig. 7 and Fig. 8 are the visualization results on the Multiome PBMC and BMMC datasets respectively. As shown in Fig. 7, the clusters obtained by CMVAE and Seurat-v4 are overall tighter, while the significant differences between CMVAE and the other models are marked by red circles. CMVAE clearly divides the NK cell population into two clusters, which implies that NK can be classified into two subtypes, and this is in fact the case, as [58] has classified NK cells in the Multiome PBMC dataset into two subtypes CD56 (bright) NK cells and CD56 (dim) NK cells. In addition, as shown in Fig. 8, CMVAE and Seurat-v4 have better discrimination for Plasma cells in the hidden variable distribution. This suggests that CMVAE makes it possible to detect more subtle differences in the hidden variable distribution by mining the correlations between RNA and ATAC, which is biologically meaningful.

To quantitatively assess the performance of the different methods, we use some common biological protection metrics the same as in [59], including Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), Average Silhouette Width (ASW) of cell type which measure the degree of retention of biological variation. The results are summarized in Table 5, where CMVAE achieved the best performance in both multi-omics datasets. Specifically, in the fully paired Multiome PBMC dataset, CMVAE performs close to Seurat-v4, while far outperforming MultiVI in the NMI and

TABLE 5
Performance comparison of cell typing. Larger values indicate better performance. The optimal and suboptimal results are in bold and underlined, respectively.

Datasets	Methods	NMI	ARI	ASW
Multiome PBMC	MOFA+ [54]	79.12	71.58	62.14
	Seurat-v4 [55]	<u>81.68</u>	<u>74.34</u>	<u>60.59</u>
	MultiVI [56]	77.68	65.37	59.48
	CMVAE	**81.85**	**75.53**	**62.80**
Multiome BMMC	MOFA+ [54]	60.63	25.86	53.40
	Seruat-v4 [55]	73.67	<u>61.19</u>	58.98
	MultiVI [56]	<u>75.10</u>	60.27	<u>59.28</u>
	CMVAE	**78.56**	**68.17**	**59.89**

ARI metrics at 4.17% and 10.16%, respectively, whereas, in the incompletely paired Multiome BMMC dataset, the CMVAE is still substantially ahead of MultiVI, while MultiVI outperforms Seurat-v4. This suggests that (i) the specific settings of the incomplete view learning method play a role in incomplete view data. (ii) The proposed complete multi-view representation learning outperforms the multi-view consistent learning in the view complete or missing case because learning complete information helps to distinguish small differences between samples.

5 Conclusion

In this paper, we put forward the complete multi-view VAE (CMVAE) to learn a complete generative latent representation under view absence. Specifically, view-invariant information mining is introduced into the inference process of latent variables, allowing the missing view information to be compensated. The variational inference process includes exploiting the intrinsic transformations between views for interconversion and keeping the view weights invariant to avoid misrepresentation of the latent variable. Benchmark experiments, time complexity and parameter sensitivity analysis, and bioinformatics applications are conducted to demonstrate the effectiveness, efficiency, robustness and practical significance of the proposed multi-view variational lower bound under the VAE framework.

Acknowledgment

This work was supported in part by the National Key Research and Development Program of China (2022YFE0112200), Science and Technology Project of Guangdong Province (2022A0505050014), the Key-Area Research and Development Program of Guangzhou City (202206030009), and the National Natural Science Foundation of China (U21A20520, 62325204, 62172112).

References

[1] Y. Li, M. Yang, and Z. Zhang. A survey of multi-view representation learning. *IEEE transactions on knowledge and data engineering*, 31(10):1863-1883, 2018.
[2] G. Andrew, R. Arora, J. Bilmes and K. Livescu. Deep canonical correlation analysis. in *International conference on machine learning*, pages 1247-1255, PMLR, 2013.
[3] W. Wang, X. Yan, H. Lee, and K. Livescu. Deep variational canonical correlation analysis. *CoRR*, abs/1610.03454, 2016.
[4] K. Andrej, and F. Li. Deep visual-semantic alignments for generating image descriptions. in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128-3137, 2015.
[5] Y. Zhao, X. You, S. Yu, C. Xu, W. Wei, X.Y. Jing, T. Zhang and D. Tao. Multi-view manifold learning with locality alignment. *Pattern recognition*, 78:154-166, 2018.
[6] S. Nitish, and S. Russ R. Multimodal learning with deep boltzmann machines. in *Advances in neural information processing systems*, 2012.
[7] C. Zhang, H. Fu, J. Wang, W. Li, X. Cao and Q. Hu. Tensorized multi-view subspace representation learning. *International journal of computer vision*, 128(8):2344-2361, 2020.
[8] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal Deep Learning. in *International conference on machine learning*, 2011.
[9] X. Jia, X.Y. Jing, X. Zhu, S. Chen, B. Du, Z. Cai, Z. He and D. Yue. Semi-supervised multi-view deep discriminant representation learning. *IEEE transactions on pattern analysis and machine intelligence*, 43(7):2496-2509, 2020.
[10] M. Hu and S. Chen. Doubly aligned incomplete multi-view clustering. in *International joint conference on artificial intelligence*, pages 2262-2268, 2018.
[11] M. Hu and S. Chen. One-pass incomplete multi-view clustering. in *Proceedings of the AAAI conference on artificial intelligence*, pages 3838-3845, 2019.
[12] C. Shang, A. Palmer, J. Sun, K. Chen, J. Lu, and J. Bi. VIGAN: Missing view imputation with generative adversarial networks. in *IEEE international conference on big data*, pages 766-775, 2017.
[13] Q. Wang, Z. Ding, Z. Tao, Q. Gao, and Y. Fu. Generative Partial Multi-View Clustering With Adaptive Fusion and Cycle Consistency. *IEEE transactions on image processing*, 30:1771-1783, 2021.
[14] L. Yuan, Y. Wang, P. M. Thompson, V. A. Narayan, and J. Ye. Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data. in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* , pages 1149-1157, 2012.
[15] T. M. Sutter, I. Daunhawer, and J. E. Vogt. Generalized multimodal elbo. in *International conference on learning representations* , 2021.
[16] S. Li, Y. Jiang, and Z. Zhou. Partial multi-view clustering. in *Proceedings of the AAAI conference on artificial intelligence*, 2014.
[17] C. Zhang, Y. Cui, Z. Han, J. T. Zhou, H. Fu, and Q. Hu. Deep partial multi-view learning. *IEEE transactions on pattern analysis and machine intelligence*, 2020.
[18] A. Li, H. Hu, P. Mirowski and M. Farajtabar. Cross-view policy learning for street navigation. reconstruction from human brain activity. in *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8100-8109, 2019.
[19] Q. Wang, H. Lian, G. Sun, Q. Gao, and L. Jiao. ICMSC: Incomplete cross-modal subspace clustering. *IEEE transactions on image processing*, 30:305-317, 2020.
[20] H. Hotelling. Relations between two sets of variates. in *Breakthroughs in statistics*, pages 162-190. Springer, 1992.
[21] S. Akaho. A kernel method for canonical correlation analysis. in *International meeting of psychometric society*, 2001.
[22] W. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep multi-view representation learning. in *International conference on machine learning*, pages 1083-1092, 2015.
[23] C. Zhang, H. Fu, Q. Hu, P. Zhu, and X. Cao. Flexible multi-view dimensionality co-reduction. *IEEE transactions on image processing*, 26(2):648-659, 2016.
[24] H. Zhao, Z. Ding, and Y. Fu. Multi-view clustering via deep matrix factorization. in *Proceedings of the AAAI conference on artificial intelligence*, 2017.
[25] J. Yin, and S. Sun. Multiview uncorrelated locality preserving projection. *IEEE transactions on neural networks and learning systems*, 31(9):3442-3455, 2019.
[26] X. Wu, Q. Chen, Y. Hu, D. Wang, X. Chang, X. Wang, and M. Zhang. Multi-View Multi-Label Learning with View-Specific Information Extraction. in *International joint conference on artificial intelligence*, pages 3884-3890, 2019.
[27] Y. Liang, D. Huang, C. Wang, and S. Y. Philip. Multi-view graph learning by joint modeling of consistency and inconsistency. *IEEE transactions on neural networks and learning systems*, 2022.
[28] W. Shao, X. Shi, and S. Y. Philip. Clustering on multiple incomplete datasets via collective kernel learning. in *IEEE 13th international conference on data mining*, pages 1181-1186, 2013.

[29] C. Xu, D. Tao, and C. Xu. Multi-view learning with incomplete views. *IEEE transactions on image processing*, 24(12):5812-5825, 2015.
[30] L. Tran, X. Liu, J. Zhou, and R. Jin. Missing modalities imputation via cascaded residual autoencoder. in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1404-1414, 2017.
[31] H. Zhao, H. Liu, and Y. Fu. Incomplete multi-modal visual data grouping. in *International joint conference on artificial intelligence*, pages 2392-2398, 2016.
[32] X. Li, M. Chen, C. Wang, and J. Lai. Refining graph structure for incomplete multi-view clustering. *IEEE transactions on neural networks and learning systems*, 2022
[33] Y. Lin, Y. Gou, X. Liu, J. Bai, J. Lv, and X. Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE transactions on pattern analysis and machine intelligence*, 2022.
[34] C. Zhang, Z. Han, H. Fu, J. T. Zhou, and Q. Hu. CPM-Nets: Cross partial multi-view networks. in *Advances in neural information processing systems*, 2019.
[35] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.
[36] M. Wu, and N. Goodman. Multimodal generative models for scalable weakly-supervised learning. in *Advances in neural information processing systems*, 2018.
[37] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771-1800, 2002.
[38] Y. Shi, B. Paige, and P. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models. in *Advances in neural information processing systems*, 2019.
[39] T. Sutter, I. Daunhawer, and J. Vogt. Multimodal generative learning utilizing jensen-shannon-divergence. in *Advances in neural information processing systems*, 2020.
[40] M. Yin, W. Huang and J. Gao. Shared generative latent representation learning for multi-view clustering. in *Proceedings of the AAAI conference on artificial intelligence*, pages 6688-6695, 2020.
[41] Y. Lin, Y. Gou, Z. Liu, B. Li, J. Lv, and X. Peng. COMPLETER: Incomplete multi-view clustering via contrastive prediction. in *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11174-11183, 2021.
[42] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. in *International conference on learning representations*, 2015.
[43] J. Winn, and N. Jojic. Locus: Learning object classes with unsupervised segmentation. in *Proceedings of the IEEE/CVF international conference on computer vision*, pages 756-764, 2005.
[44] Y. Zhang, C. Xu, H. Lu, and Y. Huang. Character identification in feature-length films using global face-name matching. *IEEE transactions on multimedia*, 11(7):1276-1288, 2009.
[45] X. Cao, C. Zhang, H. Fu, S. Liu, and H. Zhang. Diversity-induced multi-view subspace clustering. in *Proceedings of the IEEE/CVF international conference on computer vision*, pages 586-594, 2015.
[46] A. Asuncion and D. Newman. Uci machine learning repository. [http://archive.ics.uci.edu/ml], 2007.
[47] F. Li, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. in *Proceedings of the IEEE/CVF international conference on computer vision workshop*, 2004.
[48] X. Cai, H. Wang, H. Huang, and C. Ding. Joint stage recognition and anatomical annotation of drosophila gene expression patterns. *Bioinformatics*, 28(12):i16-i24, 2012.
[49] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):453-465, 2013.
[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. in *Advances in neural information processing systems*, pages 1097-1105, 2012.
[51] K. Simonyan, and A. Zisserman. Very deep convolutional networks for large-scale image recognition. in *International conference on learning representations*, 2015.
[52] J. Wen, Z. Zhang, Y. Xu, B. Zhang, L. Fei, and H. Liu. Unified embedding alignment with missing views inferring for incomplete multi-view clustering. in *Proceedings of the AAAI conference on artificial intelligence*, pages 5393-5400, 2019.
[53] M.D. Luecken, D.B. Burkhardt, R. Cannoodt, C. Lance, A. Agrawal, H. Aliee, A.T. Chen, L. Deconinck, A.M. Detweiler, A.A. Granados and S. Huynh. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. in *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (Round 2)*., 2021.
[54] R. Argelaguet, D. Arnol, D. Bredikhin, Y. Deloro, B. Velten, J.C. Marioni and O. Stegle. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome biology*, 21(1):1-17, 2020.
[55] Y. Hao, S. Hao, E. Andersen-Nissen, W.M. Mauck, S. Zheng, A. Butler, M.J. Lee, A.J. Wilk, C. Darby, M. Zager and P. Hoffman. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573-3587, 2021.
[56] T. Ashuach, M.I. Gabitto, R.V. Koodli, G.A. Saldi, M.I. Jordan and N. Yosef. Integrated analysis of multimodal single-cell data. *Nature Methods*, 20(8):1222-1231, 2023.
[57] L. McInnes, J. Healy, N. Saul, and L. Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.
[58] S. Ghazanfar, C. Guibentif and J.C. Marioni. Stabilized mosaic single-cell data integration using unshared features. *Nature Biotechnology*, 1-9, 2023.
[59] M.D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M.F. Müller, D.C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché and F.J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41-50, 2022.

Hongmin Cai (Senior Member, IEEE) is a Professor at the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. He received the B.S. and M.S. degrees in mathematics from the Harbin Institute of Technology, Harbin, China, in 2001 and 2003, respectively, and the Ph.D. degree in applied mathematics from Hong Kong University in 2007. From 2005 to 2006, he was a Research Assistant with the Center of Bioinformatics, Harvard University, and Section for Biomedical Image Analysis, University of Pennsylvania. His areas of research interests include biomedical image processing and omics data integration.

Weitian Huang received the MS degree in Control Science and Engineering from the School of Automation, Guangdong University of Technology, Guangzhou, China, in 2020. He is currently working toward the PhD degree from the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. His research interests include multi-view learning, deep generative model, clustering, and bioinformatics.

Sirui Yang received the bachelor's degree in computer science from the China University of Mining Technology, Xuzhou, China. He is currently working toward the master's degree from the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. His current research interests include data mining and multi-view clustering.

Siqi Ding received the bachelor's degree in computer science from the Jiangnan University, Wuxi, China. She is currently working toward the master's degree from the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. Her current research interests include multi-view learning and clustering.

Yue Zhang received the Ph.D. degree in computer science from Hong Kong Baptist University, Hong Kong SAR, China, in 2017. She is an Associate Professor with the School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China. Her research interests include bioinformatics and big data mining.

Bin Hu (Fellow, IEEE) received Ph.D. degree in computer science from Institute of Computing Technology, Chinese Academy of Science, China in 1998. Since 2008, he has been a professor and the Dean of School of Information Science and Engineering, Lanzhou University, China. He had been also guest professorship in ETH Zurich, Switzerland till 2011. His research interests include Pervasive Computing, Computational Psychophysiology, and Data Modeling.

Fa Zhang received the PhD degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China. He is a chair professor at ICT, CAS. His current research interests include bioinformatics, biomedical image processing, and high-performance computing.

Yiu-ming Cheung (Fellow, IEEE) received the PhD degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong SAR, China. He is currently a chair professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His current research interests include machine learning, pattern recognition, and visual computing. He is the founding chairman of the Computational Intelligence Chapter of the IEEE Hong Kong Section. He is the Editor-in-Chief of IEEE Transactions on Emerging Topics in Computational Intelligence. Also, he serves as an Associate Editor for IEEE Transactions on Cybernetics, IEEE Transactions on Cognitive and Developmental Systems, Pattern Recognition, Neurocomputing, to name a few. He is a Fellow of the IEEE, AAAS, IET and BCS.