Cluster Induced Generative Incomplete Image-Text Clustering (CIGIT-C)

Xiaoming Su  
College of data science and Application, Inner Mongolia University of Technology, Hohhot, China  
xiaoming_su@foxmail.com

Dongjin Guo  
College of data science and Application, Inner Mongolia University of Technology, Hohhot, China  
20201100114@imut.edu.cn

Rui Su  
College of Electric Power, Inner Mongolia University of Technology, Hohhot, China; Inner Mongolia Power(Group)Co,Ltd, Hohhot, China  
20190000015@imut.edu.cn

Pei Zhiyong  
College of energy and transportation engineering, Inner Mongolia Agricultural University, Hohhot, China  
peizhiyong@imau.edu.cn

ABSTRACT
The target of incomplete image-text clustering (IITC) is to find correct clusters by integrating complementary and consistent information of multi-modalities for these unpaired heterogeneous samples. Although a series of methods have been proposed to address this issue, the following problems still exist: 1) Most existing methods hardly consider the distinct gap between heterogeneous feature domains. 2) For missing data, the representations generated by existing methods are rarely guaranteed to suit clustering tasks. 3) Existing methods do not tap into the latent connections between inter and intra modalities. In this paper, we propose a Clustering Induced Generative Incomplete Image-Text Clustering (CIGIT-C) network to address the challenges above. More specifically, we first use modality-specific encoders to map original features to more distinctive subspaces. The latent connections between intra and inter-modalities are thoroughly explored by using the adversarial generating network to produce one modality conditional on the other modality. Finally, we update the corresponding modality-specific encoders using two KL divergence losses. Experiment results on public image-text datasets demonstrated that the suggested method outperforms the state-of-the-art baseline methods on 2 benchmark datasets with missing rate of 50% and 70%.

CCS CONCEPTS
• Computing methodologies → Artificial intelligence; Computer vision; Computer vision representations.

KEYWORDS
Unsupervised Multi-modal Representations, Incomplete Image-text Clustering, Adversarial generating network

1 INTRODUCTION
The two primary ways of displaying information in daily life are visual media and natural language. Therefore, as a fundamental research topic in machine learning and data science communities, ITC has drawn an increasing amount of interest in recent years. It is advantageous for a variety of applications, including vehicle detection [1], clinical medicine [2], marketing research, crowd sourced mobile testing [3] and recommendation system [4]. However, in real-life, some instances may lack paired instances of different modalities. For instance, on Twitter, tweets can contain texts or images, but not all contain paired modalities. The missing modalities not only lead to information loss, but also increase the difficulties of excavating complementary information. In general, most present IITC algorithms rely heavily on the idea of incomplete multi-view clustering. Over the past few years, several efforts have been dedicated to addressing IITC, with various imputation strategies (i.e., zero or mean-based methods, common space learning-based methods, and generative methods) to fill in the missing samples. As shown in Fig. 1(a), these methods [5] adopt zero or mean values to pad missing data first and then design specific machine learning techniques to perform multi-view clustering. These simply padded data are very different from the real data and can reduce the performance of the final clustering. To alleviate this challenge, a common latent space learning-based method has been introduced, as shown in Fig. 1(b). The goal of the method is to learn a common latent representation using complete and aligned modalities and then use the common representation to complete the representation of the missing modalities, such as the kernel [6] [7] method, the matrix [8] [9] method and the contrastive learning [10] method. However, these methods cannot explicitly compensate for the missing data in each modality. To solve this issue, generative-based(GAN [11])methods(see Fig. 1(c) and 1(d)) are proposed. As shown in the Fig. 1(c), the third imputation strategy for missing values is based on Cycle GAN [12], such as TPIT-C [13].
Figure 1: Four imputation strategies for missing values. (a) Zero or mean-based methods. (b) Common space learning-based method. (c) Cycle GAN-based method. (d) Ours proposed CIGIT-C. Because of their heterogeneity, the solid triangles and circles in the figure represent the image and text modalities. Circles and triangles of the same colour represent the same instance. The graph drawn in the dashed box is a representation of the missing modalities.

Although the aforementioned methods provide some schemes to address the IITC problem, it still suffer from the following issues:

1) The majority of previously incomplete multi-view clustering techniques extract distinct view features from various visual datasets that are fundamentally homogeneous using various feature description techniques (e.g., SIFT, LBP, HOG [14]).
2) The inaccurate imputation or padding of missing data has the potential to negatively impact cluster performance. In addition, the missing data processing tends to lean towards the generation task rather than the clustering task. Therefore, the performance in completing and inverting the missing value is poor.
3) Existing methods ignore the latent connections within and between the modalities.

To address the above issues, we propose a novel model, named CIGIT-C, which first maps the original features of the image and text into their respective subspace by their specific encoders. The different modalities provide distinctive representations at the cluster level. This step takes into account the intra-modal connections.

Second, inspired by conditional GAN [15], as shown in Fig. 1(d), the two generators use the knowledge of the cluster label as conditional information provided by their respective subspace to standardize the process of representation generation, which is beneficial for the clustering results. Clustered adversarial networks can not only generate representations to compensate for missing data. Moreover, these generated representations are beneficial for clustering tasks. The generators thoroughly learn the knowledge of clustering-level across modalities, and fully exploiting the complementary between modalities. KL clustering loss is used to update the encoder, which attempts to make the distribution of the learned representations consistent and compact. The main contributions of our approach are listed below:

(a) We have argued theoretically (in Section I) and experimentally (in Section IV) that existing incomplete multi-view methods do not solve the IITC task well. Such a theoretical view is remarkably different from existing works, which treat image-text as homogeneous data.

(b) For the processing of missing data, we are driven by clustering tasks rather than generating tasks, which are ignored in the existing IITC methods.

(c) We fully explore the potential relationships within and between modalities. Extensive experiments demonstrate that our method achieves superior clustering performance compared to state-of-the-art methods.

2 RELATED WORK

2.1 Incomplete Multi-view Clustering

Existing incomplete multi-view clustering methods can be divided into traditional methods [16] [17] [18] [19] and deep learning based methods [24] [25] [26] [27]. For example, DAIMC [16] establishes a consensus basis matrix with the help of $\ell_2,1$-norm. IMG [17] employs the graph Laplacian to regularize the latent subspace of each data view. Hu et al. [18] propose an efficient and effective method to handle large-scale incomplete multi-view clustering problems using regularized matrix decomposition and weighted matrix factorization, taking full account of the missing information of the instances. The GAN model is intended to discover the connection between various modalities. The model from references [24] was used to complete the missing data. Cycle GAN [25] uses generative methods and their inverse directions to achieve unpaired image style transformation.

2.2 Knowledge Distillation

Knowledge distillation [28] is a compression technique in which a compact model (student) is trained to reproduce the behavior of a larger model (teacher) or set of models. Distilling models [29] [30] convey extra information beyond the traditional supervised learning target by mimicking the teacher’s class probability or feature representation. Recently, in order to address the inconsistency between training and inference caused by the introduction of dropout, Liang et al. [31] introduce a simple consistent training strategy that normalizes dropout. As some modalities in the IITC task might be inherent in low quality or inconsistent cross-modality clustering prediction, they will negatively affect the fusion process, in turn, the clustering result. Motivated by these, we design a loss function that minimizes KL-divergence between the clustering prediction scores of the respective modalities and fused representations clustering probability.
Clustering-Induced Generative Incomplete Image-Text Clustering (CIGIT-C) ASSE 2023, October 27–29, 2023, Aizu-Wakamatsu City, Japan

Figure 2: The framework of our proposed Clustering-Induced Generative Incomplete Image-Text Clustering (CIGIT-C).

Table 1: NOTATIONS AND DESCRIPTIONS

| Notations | Descriptions |
|-----------|--------------|
| X\textsuperscript{img}, X\textsuperscript{txt} | Image/Text original features. |
| ̃X\textsuperscript{img}, ̃X\textsuperscript{txt} | Text/Image features of data missing modalities. |
| Z\textsuperscript{img}, Z\textsuperscript{txt} | Subspace representation of images/text. |
| E\textsuperscript{img}, E\textsuperscript{txt} | Encoders of images/text representation |
| z\textsuperscript{fake}\textsuperscript{img}, z\textsuperscript{fake}\textsuperscript{txt} | Generated fake images/text representations. |
| Z\textsuperscript{fusion} | Fusion representation |

3 CIGIT-C: THE PROPOSED MODEL

In this section, we first introduce the notations and descriptions. After clarifying this, we describe each component and the loss function in detail. Then, we provide a method to optimize the CIGIT-C model. Finally, we integrate the above to provide a discussion of the CIGIT-C model.

3.1 Notations and Descriptions

Define Given a dataset \( D = \{X_{\text{img}}, X_{\text{txt}}\} \) of \( n \) instances, where \( X_{\text{img}}, X_{\text{txt}} \) denote the examples presented in both modalities, the visual modality only, and the text modality only, respectively. For clarity, we denote variables originated from images and texts by superscript \text{img} and \text{text}, respectively. The frequently used notations and their descriptions are listed in Table 1.

3.2 Model Overview

The original features of the image and text are first mapped to a more distinctive subspace via encoders \( E_{\text{img}} \) and \( E_{\text{txt}} \) respectively. More distinctive representations were found by subspace clustering. The two generators generate a clustering task-driven representation based on the clustering information provided by the other subspace. The generative mechanism fully uses the characteristics and cluster distribution of \( Z_{\text{img}} \) and \( Z_{\text{txt}} \). The generated representation is fused with the original representation for clustering. Furthermore, at the same time, the aligned clustering loss is designed to obtain a better modality-specific encoder.

3.3 Modality-Specific Encoders

The role of the modality-specific encoders \( E_{\text{img}} \) and \( E_{\text{txt}} \) to map the original features into a more distinctive subspace.

\[
Z_{\text{ fus}} = \sum_{t=1}^{n} (1 - \beta)E_{\text{img}}(X_{\text{img}}^t) + \beta E_{\text{txt}}(X_{\text{txt}}^t) \quad (1)
\]

\( E_{\text{img}}(x_{\text{img}}, y_{\text{txt}}) \) and \( E_{\text{txt}}(x_{\text{txt}}, y_{\text{txt}}) \) in Eq (1) represent the encoders for images and text, respectively, where \( \theta \) is the parameter that can be learned by back-propagation. \( z_{\text{img}}^t = E_{\text{img}}(x_{\text{img}}^t) \) and \( z_{\text{txt}}^t = E_{\text{txt}}(x_{\text{txt}}^t) \) represent the subspace features, while \( Z_{\text{ fus}}^t \) is the subspace features of the fused image and text, where \( \beta \) is a weighting factor to balance the ratio of text and visual modality. Moreover, in order to make the projected samples more distinctive across modalities, the Student’s t-distribution is utilized, where the goal of Student’s t-distribution is to make the projected representations closer to the samples of the same clustering center than it is to any other cluster center. As follow Eq (2), \( \delta \) represents the degrees of freedom of the Student’s t-distribution. Following Xie et al. [22], we set \( \delta \) to 1. The clustering centroid of the K-th
cluster in the subspace for the m-th modality is denoted by $\mu_k^m$. The centroids are initialized by K-means and updated by Stochastic Gradient Descent (SGD). $z_{nk}^m$ represents learning the subspace representation learned through the encoder, where $m=1$ represents the visual modality, $m=2$, represents the text modality, and $m=3$ represents the fusion modality. Represents the probability of assigning instance $n$ to group $K$ of the modality (i.e., soft assignment).

$$q_{nk}^m = \frac{(1 + \left\| z_{nk}^m - \mu_k^m \right\|_2^2)^{-\frac{\beta_k}{2}}}{\sum_{k'}(1 + \left\| z_{nk}^m - \mu_k^m \right\|_2^2)^{-\frac{\beta_k}{2}}}$$  

(2)

To enhance cluster compactness, we prioritize data points assigned with high confidence by calculating the emphasised target distribution $p_{nk}$ as follows:

$$p_{nk} = \frac{q_{nk}^2}{\sum_k q_{nk}^2}$$  

(3)

where $n = \sum q_{nk}$ denotes soft cluster frequencies. Before normalising it by frequency per cluster, $f_k$ first squares $q_{nk}$ and after normalising it by frequency per cluster, it first squares $q_{nk}^2$. Thus, emphasis will be placed on high probability. Thus the modality-specific encoder is updated by the following distortion loss $L_{Em}$, as follows:

$$L_{Em} = KL(P^m || Q^m) + \alpha KL(P^3 || Q^3)$$

$$= \sum_{n} p_{nk}^m \log \frac{p_{nk}^m}{q_{nk}^m} + \alpha \sum_{n} p_{nk}^3 \log \frac{p_{nk}^3}{q_{nk}^3}$$  

(4)

where $m=1$ and $m=2$ correspond to the losses of encoders $E_{img}$ and $E_{txt}$, respectively, and $\alpha$ refers to a trade-off parameter. The objective $L_{Em}$ is to match the soft assignment $p_{nk}$ with the desired distribution $q_{nk}$. Loss $L_{Em}$ is designed to resolve inconsistencies in cluster predictions for different specific modalities. Loss $L_{Em}$ forces the distribution of fused features to be consistent with that of different modalities. Specifically, minimize the KL-divergence of fusion feature with particular modality image and text.

### 3.4 Subspace Conditional Clustering GAN

Subspace conditional clustering GAN module contains two generators, $G_{12}$ and $G_{21}$, and their corresponding discriminators, $D_1$ and $D_2$, which are trained in the inverse direction. The generator is trained to generate samples while the discriminator attempts to differentiate the samples. Both networks are compelled to improve their capabilities through a competition strategy. Since $G_{12}, D_1$ and $G_{21}, D_2$ have symmetrical positions and identical objective equations. So here, we only discuss $G_{12}$ and $D_1$.

$$L_{G_{12}d} = -E_{z-p_z(z)} \log(1 - D_1(G_{12}(z[E^{img}(X_{t}^{img})])))$$  

(5)

where $z$ refers to noise, subspace conditional clustering GAN differs from traditional GAN in focusing on clustering rather than generation tasks. $z$ is a prior sampling method consisting of a cascade of normal random variables and one-hot label noise. Since the subspaces $Z^{img}$ and $Z^{txt}$ are modified when the encoders $E^{img}$ and $E^{txt}$ are tuned, it is challenging to achieve stable generative outputs directly. Therefore, we introduce a similarity constraint that pulls produced samples and real samples toward a similar subspace. The objective phrase is depicted below:

$$L_{G_{12}2} = E_{z-p_z(z)}(\|G_{12}(z[E^{img}(X_{t}^{img})]) - E^{txt}(X^{txt})\|^2_F)$$  

(6)

The final overall objective $L_{G_{12}}$ is obtained by combining Equations 5 and 6, as shown in Eq.7 below, where $\mu$ is used to balance discriminator loss and similarity loss.

$$L_{G_{12}} = L_{G_{12}d} + \mu L_{G_{12}2}$$  

(7)

The function of $D_1$ is to distinguish the generated samples as much as possible from the real samples in subspace $Z^{txt}$. $D_1$ is the text modality that balances the weight between the real representation and the generated representation. $m=1$ represents the visual modality, while $m=2$ represents the text modality. Finally, we define the overall loss function of CIGIT-C as:

$$L_{total} = \min_{E_{m}G_{12},G_{21},D_{1},D_{2}} L_{Em} + L_{G_{12}1} + L_{G_{21}} + L_{D_{1}} + L_{D_{2}}$$  

(10)

For a clear understanding, we summarize the optimization steps of the proposed method CIGIT-C in Algorithm 1

**Algorithm 1 Clustering-Induced Generative Incomplete Image-Text Clustering (CIGIT-C) Algorithm.**

**Require:** $D = \{X^{img},X^{txt},\hat{X}^{img},\hat{X}^{txt}\}$, where some samples either lack image or text modality; learning rate $\omega$, hyper-parameter $\beta, \alpha, \eta_1, \eta_2$; iteration number $I$; cluster number $K$

**Ensure:** The optimal parameters $\Phi_{G_{12}}, \Phi_{G_{21}}, \Phi_{D1}$ and $\Phi_{D2}$: clustering results $R$;

1. Map original features $\{X^{img},X^{txt}\}$ to more distinctive subspaces $\{Z^{img}, Z^{txt}\}$. Initialize the parameters of CIGIT-C with Xavier initializer. Calculate fusion features $Z^{fusion}$ and clustering centres $\mu_k^{m}$.

2. while $\varepsilon$ ≤ MaxIter do

3. Train the encoders with $L_{Em}$ according to Eq.4. $\varepsilon_{m}=1,2$

4. Train generators $G_{12}$ and $G_{21}$ using loss functions $L_{G_{12}}$ and $L_{G_{21}}$ according to Eq.7

5. Train discriminators $D_1$ and $D_2$ using loss functions $L_{D_1}$ and $L_{D_2}$ according to Eq.8

6. The fusion features are updated according to Eq. 9, and $\mu_k^m$ is recalculated $\varepsilon_{m}=1,2,3,4,5,6,7,8,9$

7. end while

8. The final updated fusion representation $Z^{fusion}$ is obtained, and $\mu_k^m$ and $q_{nk}^m$ are computed in turn by using Eq.2 and 3, respectively.

9. Obtain the clustering result $R$;

10. return $R$;
We choose two classical and widely used clustering metrics to evaluate the performance of our models. For the task of multi-modal clustering, we adopt a pre-trained BERT to extract 768-dimensional word embeddings (BoW) vector with the TF-IDF weighting approach to get 1000-dimensional features. Similar to the BERT paper [24], we take the embedding associated with [CLS] to represent the whole sentence. For the text samples of NUS-WIDE-10K, we adopt the widely-used bag-of-words (BoW) vector with the TF-IDF weighting approach to get 1000-dimensional features. The detailed statistics of the three datasets are summarized in Table II.

### 4.1 Data and features

We use state-of-the-art methods to demonstrate the effectiveness of our model CIGIT-C. We use the same image and text features for a fair comparison. For the image samples of Wikipedia and NUS-WIDE-10K, we first resize it into 224×224 then employ the VGGNet-19 [23] model pre-trained on the ImageNet to output 4096-dimensional features from its fc7 layer. For the text samples of Wikipedia, we adopt a pre-trained BERT to extract 768-dimensional text features. Similar to the BERT paper [24], we take the embedding associated with [CLS] to represent the whole sentence. For the text samples of NUS-WIDE-10K, we adopt the widely-used bag-of-words (BoW) vector with the TF-IDF weighting approach to get 1000-dimensional features. The detailed statistics of the three datasets are summarized in Table II.

### 4.2 Data construction

To demonstrate the effectiveness of the CIGIT-C model in processing missing image-text data. We set different missing rates $p \in \{10\%, 30\%, 50\%, 70\%\}$. The missing rate $p$ is defined as $p = (n - m)/n$, where $m$ is the number of complete examples, and $n$ is the number of the whole dataset. We randomly delete some instances as missing data in a modality by the missing rate $p$. A high missing rate is positively correlated with task complexity.

### 4.3 Evaluation Metric

We choose two classical and widely used clustering metrics to measure the consistency of cluster assignments and ground-truth memberships:

- **Clustering accuracy (ACC)** maps the one-to-one learned clusters to the ground truth classes by the Hungarian algorithm and measures the classification accuracy;

- **Normalized mutual information (NMI)** quantifies the labelling consistency by the normalized MI between the predicted and ground truth labels of all samples.

The real number of clusters $k$ is set as the true number of classes for all data sets under the assumption that it is known. We initialize 20 times and choose the best answer for each approach using the k-means algorithm to produce clustering assignments. In order to prevent extreme occurrences, we run each method 10 times and present the average results.

### 4.4 Experimental Results

Tables III to V present the details of the network architectures in these two training modules, respectively. Take the Wikipedia dataset, $E^{txt}$ and $E^{img}$ are three-layer networks in our implementation. We set the dimension of $E^{txt}$ to 4096-1024-256, where 4096 is the feature dimension of the visual modality, 1024 denotes the hidden layer dimension, and 256 denotes the subspace representation dimension. We set the dimension of the $E^{img}$ to 728-256-128. $G^{12}$ and $G^{21}$ are three-layer neural networks with a batch normalization layer to normalize the input vector and stabilize the training procedure. The proposed $D^1$ and $D^2$ are mainly made up of a fully connected layer with ReLU [32] activation, a mini-batch layer that can increase the diversity of fake representations, a sigmoid function which outputs the fake-real possibility of input representations.
### Table 3: THE ARCHITECTURE OF ENCODERS IN CIGIT-C

| Datasets       | $E^{\text{img}}$                                                                 | $E^{\text{txt}}$                                                                 |
|----------------|---------------------------------------------------------------------------------|---------------------------------------------------------------------------------|
| Wikipedia      | Input (size=(bathsize,4096))                                                     | Input (size=(bathsize,768))                                                     |
|                | Dense (LeakyRelu, size=2048)                                                     | Dense (LeakyRelu, size=256)                                                     |
|                | Dense (LeakyRelu, size=1024)                                                     | Dense (LeakyRelu, size=256)                                                     |
|                | Dense (LeakyRelu, size=256)                                                     | Dense (LeakyRelu, size=128)                                                     |
| NUS-WIDE-10K   | Input (size=(bathsize,4096))                                                     | Input (size=(bathsize,1000))                                                    |
|                | Dense (LeakyRelu, size=2048)                                                     | Dense (LeakyRelu, size=256)                                                     |
|                | Dense (LeakyRelu, size=1024)                                                     | Dense (LeakyRelu, size=256)                                                     |
|                | Dense (LeakyRelu, size=256)                                                     | Dense (LeakyRelu, size=128)                                                     |
| BDGP           | Input (size=(bathsize,1000))                                                     | Input (size=(bathsize,79))                                                      |
|                | Dense (LeakyRelu, size=256)                                                     | Dense (LeakyRelu, size=79)                                                      |
|                | Dense (LeakyRelu, size=256)                                                     | Dense (LeakyRelu, size=79)                                                      |
|                | Dense (LeakyRelu, size=128)                                                     | Dense (LeakyRelu, size=64)                                                      |

### Table 4: THE ARCHITECTURE OF GENERATORS IN CIGIT-C

| Datasets       | $G_{12}$                                                                 | $G_{21}$                                                                 |
|----------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------|
| Wikipedia      | Input (bathsize,4096+noise-dim=4396)                                     | Input (bathsize,768+noise-dim=1068)                                     |
|                | Dense (BatchNorm1d, LeakyRelu, size=1024)                                | Dense (BatchNorm1d, LeakyRelu, size=2048)                                |
|                | Dense (BatchNorm1d, LeakyRelu, size=1024)                                | Dense (BatchNorm1d, LeakyRelu, size=4096)                                |
| NUS-WIDE-10K   | Input (size=(bathsize,4096+noise-dim=4396))                             | Input (size=(bathsize,1000+noise-dim=1300))                             |
|                | Dense (BatchNorm1d, LeakyRelu, size=1024)                                | Dense (BatchNorm1d, LeakyRelu, size=2048)                                |
|                | Dense (BatchNorm1d, LeakyRelu, size=1024)                                | Dense (BatchNorm1d, LeakyRelu, size=2048)                                |
|                | Dense (BatchNorm1d, LeakyRelu, size=1000)                                | Dense (BatchNorm1d, LeakyRelu, size=4096)                                |
| BDGP           | Input (size=(bathsize,1000+noise-dim=1300))                             | Input (size=(bathsize,76+noise-dim=379))                                 |
|                | Dense (BatchNorm1d, LeakyRelu, size=256)                                 | Dense (BatchNorm1d, LeakyRelu, size=512)                                  |
|                | Dense (BatchNorm1d, LeakyRelu, size=128)                                 | Dense (BatchNorm1d, LeakyRelu, size=512)                                  |
|                | Dense (BatchNorm1d, LeakyRelu, size=79)                                  | Dense (BatchNorm1d, LeakyRelu, size=1000)                                 |

### Table 5: THE ARCHITECTURE OF DISCRIMINATORS IN CIGIT-C

| Datasets       | $D_{1}$                                                                 | $D_{2}$                                                                 |
|----------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------|
| Wikipedia      | Input (bathsize, $z^{\text{txt}}_{\text{fake}} + q_{nk}^{2} = 768+10=778$) | Dense (LeakyRelu, size=512)                                             |
|                | Dense (LeakyRelu, size=512)                                              | Mini-batch (LeakyRelu, size=512)                                         |
|                | Dense (sigmoid, size=1)                                                  | Dense (sigmoid, size=1)                                                  |
| NUS-WIDE-10K   | Input (bathsize, $z^{\text{img}}_{\text{fake}} + q_{nk}^{2} = 768+10=778$) | Input (bathsize, $z^{\text{img}}_{\text{fake}} + q_{nk}^{2} = 4096+10=4160$) |
|                | Dense (LeakyRelu, size=512)                                              | Dense (LeakyRelu, size=1024)                                             |
|                | Dense (sigmoid, size=1)                                                  | Dense (sigmoid, size=1)                                                  |
| BDGP           | Input (bathsize, $z^{\text{img}}_{\text{fake}} + q_{nk}^{2} = 79+10=89$) | Dense (LeakyRelu, size=64)                                              |
|                | Dense (LeakyRelu, size=64)                                               | Mini-batch(LeakyRelu, size=512)                                          |
|                | Dense (sigmoid, size=1)                                                  | Dense (sigmoid, size=1)                                                  |

5) For BDGP dataset, the result of certain methods (like OPIMCC and COMPLETER) is poor. The reason is 5.1) As the missing rate rises, complementary information between multi-modalities becomes scarce, and all methods accuracy inevitably decline. 5.2) The case of high missing rates, it is difficult to find connections between modalities. 3.3) The CDIMC-net padding strategy depends on the estimation of the data distribution. As a result, the cumulative error increases when the missing rate is high.
Clustering-Induced Generative Incomplete Image-Text Clustering (CIGIT-C)

Table 6: CLUSTERING RESULTS (ACC AND NMI) OF DIFFERENT METHODS ON THE WIKIPEDIA DATASET WITH DIFFERENT MISSING RATE p%

| Method       | 10%  | 30%  | 50%  | 70%  | 10%  | 30%  | 50%  | 70%  |
|--------------|------|------|------|------|------|------|------|------|
| BestSM       | 44.79| 40.70| 33.45| 23.32| 48.55| 39.51| 31.80| 22.31|
| DAIMC [16]   | 55.05| 43.70| 31.96| 18.99| 45.18| 28.43| 16.84| 9.86 |
| UEAF [27]    | 53.03| 43.53| 33.79| 25.12| 49.83| 38.12| 26.64| 18.59|
| OPIMCC [11]  | 45.15| 28.28| 16.92| 9.010| 54.43| 39.92| 33.63| 25.75|
| PIC [28]     | 45.60| 28.45| 27.43| 23.18| 36.88| 33.24| 28.75| 21.16|
| CDIMC-net [4]| 31.28| 28.45| 27.43| 23.18| 36.88| 33.24| 28.75| 21.16|
| CASENN [3]   | 57.46| 50.52| 46.68| 42.22| 56.83| 42.84| 39.92| 31.79|

Table 7: CLUSTERING RESULTS (ACC AND NMI) OF DIFFERENT METHODS ON THE NUS-WIDE-10K DATASET WITH DIFFERENT MISSING RATE p%

| Method       | 10%  | 30%  | 50%  | 70%  | 10%  | 30%  | 50%  | 70%  |
|--------------|------|------|------|------|------|------|------|------|
| BestSM       | 50.82| 47.14| 42.49| 32.88| 57.71| 48.76| 40.41| 33.25|
| DAIMC [16]   | 61.84| 51.06| 40.91| 32.18| 56.72| 45.54| 34.12| 25.42|
| UEAF [27]    | 51.90| 35.91| 23.67| 15.95| 55.60| 45.72| 40.76| 33.34|
| OPIMCC [11]  | 51.52| 47.09| 36.73| 32.91| 40.41| 34.91| 29.94| 17.47|
| PIC [28]     | 59.98| 58.41| 51.28| 47.94| 54.46| 52.82| 49.68| 34.23|
| CDIMC-net [4]| 62.39| 56.34| 50.74| 46.51| 62.63| 52.81| 47.97| 31.24|
| CASENN [3]   | 48.66| 43.79| 42.88| 39.67| 40.66| 36.66| 39.51| 33.17|
| LHGN [25]    | 48.95| 43.91| 42.43| 39.47| 41.15| 37.12| 38.95| 32.89|
| CIGIT-C      | 63.26| 57.44| 52.48| 48.39| 63.56| 54.93| 49.65| 38.73|

Similarity of the data is the primary variable influencing clustering performance. The critical component is still the multi-modal learning of features since similarity depends on characteristics. The low quality of the modal features in the BDGP dataset has a direct impact on the performance of the method.

6) DAIMC and OPIMCC are based on matrix factorization. However, their clustering performance is limited because they mainly make use of some regularization and add some constraints to the new representation but fail to compensate for the missing data in each view explicitly.

7) UEAF and PIC explore intra-modality connections and structure for multi-view clustering yet neglect to explore the underlying relationships and structure between modalities. Therefore, valid clustering results can not be obtained.

4.5 Parameter Analysis and Ablation Study

This section examines the CIGIT-C on the Wikipedia, NUS-WIDE-10K, and BDGP datasets from the perspective of parameter analysis and ablation studies. Our method contains four user-defined parameters, weighting factor $\beta$ to balance the ratio of image modality to text modality and weighting factor $\alpha$ to control KL loss. $\eta_m$ is the generated and true representation weighting coefficient for the m-th order modality. $\eta_1$ is for image modality, and $\eta_2$ is for text modality.

The $\eta_1$ and $\eta_2$ parameters are tuned in the range same as $\alpha$. The best results are achieved on the Wikipedia dataset at a missing rate of 0.5 with parameters both taking the value of 0.1. The parametric analysis of CIGIT-C on the NUS-WIDE-10K and BDGP datasets is analyzed using the same methods. Due to space constraints, we will not continue our conversation here.

To demonstrate the importance of the Clustering GAN module and KL loss in CIGIT-C, we performed ablation experiments with a miss rate of 0.5. As shown in Figure 4, where “None Clustering GAN” means that the conditional clustering GAN does not participate in the CIGIT-C work, similarly, “None KL-Divergence losses” means that CIGIT-C contains only the clustered GAN.

We can draw the following conclusions:

1) It can be seen that when both clustering GAN and KL-loss exist, CIGIT-C performs the best, indicating that modules and loss functions in CIGIT-C complement each other and promote each other.

2) The CIGIT-C model does not improve much when only KL-loss is available. Even with the consistency constraint, i.e. the
Table 8: CLUSTERING RESULTS (ACC AND NMI) OF DIFFERENT METHODS ON THE BDGP DATASET WITH DIFFERENT MISSING RATE p%

| Method       | Accuracy | NMI  |
|--------------|----------|------|
|              | 10%      | 30%  | 50%  | 70%  | 10%      | 30%  | 50%  | 70%  |
| BestSM       | 47.52    | 41.44 | 34.47 | 27.95 | 35.74    | 25.20 | 16.39 | 9.29 |
| DAIMC [16]   | 74.76    | 62.88 | 52.45 | 35.82 | 55.64    | 47.87 | 28.33 | 9.17 |
| UEAF [27]    | 76.56    | 57.13 | 50.41 | 36.84 | 65.13    | 44.10 | 33.13 | 12.60|
| OPIMCC [11]  | 71.69    | 56.44 | 55.17 | 35.82 | 61.77    | 41.47 | 25.94 | 8.35 |
| PIC [28]     | 68.26    | 63.05 | 53.23 | 44.31 | 66.70    | 55.77 | 27.78 | 11.43|
| CDIMC-net [4] | 87.50  | 76.99 | 60.47 | 40.98 | 77.24    | 57.09 | 35.64 | 14.77|
| CASENN [3]   | 79.92    | 70.96 | 59.25 | 46.93 | 72.21    | 48.23 | 29.77 | 12.45|
| COMPLETER [13]| 59.60   | 55.20 | 54.10 | 52.90 | 52.80    | 51.10 | 33.04 | 14.77|
| LHGN [25]    | 60.59    | 55.46 | 56.98 | 52.87 | 53.48    | 51.65 | 33.23 | 14.79|
| CIGIT-C      | 78.12    | 72.14 | 62.74 | 51.82 | 68.41    | 51.76 | 36.11 | 16.78|

Figure 3: Accuracy (%) performance with different $\beta$ and $\alpha$ when the missing rate $p$ is 0.5.

Figure 4: Effectiveness of clustering GAN, KL-divergence losses, when the missing rate $p$ is 0.5.

different and fused modalities predict the clustering knowledge consistently. However, the performance of CIGIT-C inevitably decreases because no treatment is given to the missing values.

3) When only clustering GAN is available, although clustering GAN serves to increase the diversity of cross-modal representations and solve the problem of missing modalities, it lacks consideration of the inconsistencies arising from inter-modal clustering predictions.

4.6 Qualitative Analysis

In order to verify the effectiveness of clustering GAN in CIGIT-C, we visualize the distribution of real and generated representations of samples in $Z_{img}$ and $Z_{txt}$ by t-SNE [33]. The results are shown in Figure. 5, which illustrates that the true and generated representations belonging to the same cluster category are close to each other and vice versa. The visualization in Figure. 6 demonstrates the advantages of our proposed CIGIT-C in terms of clustering results.

In contrast to CGIG-C, COMPLETER [13] incorrectly clusters text with the semantic category art and media images together. CDIMC-net [4] incorrectly clusters text with the semantic category music and sports images together.

5 CONCLUSIONS

We demonstrate theoretically and experimentally that the existing incomplete multi-view clustering method cannot solve the IITC task well, which is not considered in the existing works. We propose a clustering Gan to ensure that the populated data benefits the
Figure 5: t-SNE [33] visualization results of the real and the generated test sample representations in $Z^{\text{img}}$ and $Z^{\text{txt}}$ respectively.

Figure 6: The SOTA method and the CIGIG-C method visualise the clustering results. The bottom white clustering result in the figure is correct, with both images and text belonging to the same semantic category. The categories are art and music from top to bottom, while the yellow box shows the wrong clustering results.

classification task. We also add consistency constraints by introducing the KL divergence loss function to reduce the inconsistency caused by different modal predictions.

ACKNOWLEDGMENTS

This work was supported by the Key Science and Technology Special Program of inner Mongolia Autonomous Region under Grant 2021ZD0015.

REFERENCES

[1] A. Gomaa, M. M. Abdelwahab, M. Abo-Zahhad, T. Minematsu, and R.-i. Taniguchi, 2019.Robust vehicle detection and counting algorithm employing a convolution neural network and optical flow, Sensors, vol. 19, no. 20, p. 4588.
[2] E. Horne, H. Tibble, A. Shekht, A. Tisanas et al., 2020.Challenges of clustering multimodal clinical data: review of applications in asthma subtyping, JMIR Medical informatics, vol. 8, no. 5, p. e16552.
[3] B. Li, J. Li, and X. Ou, 2021.Hybrid recommendation algorithm of crossborder e-commerce items based on artificial intelligence and multiview collaborative fusion, Neural Computing and Applications, pp. 1–10.
[4] T. D. Do, K. Kim, H. Park, and H.-J. Yang, 2020.Image and encoded text fusion for deep multi-modal clustering, in The 9th International Conference on Smart Media and Applications, pp. 308–312.
[5] J. Wen, Z. Zhang, Z. Zhang, L. Zhu, L. Fei, B. Zhang, and Y. Xu, 2021 Unified tensor framework for incomplete multi-view clustering and missing-view inferring, in Proceedings of the AAAI conference on artificial intelligence, vol. 35, no. 11, pp. 10 273–10 281.
[6] X. Liu, X. Zhu, M. Li, L. Wang, C. Tang, J. Yin, D. Shen, H. Wang, and W. Gao, 2018.Late fusion incomplete multi-view clustering, IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 10, pp. 2418–2423, 2018.
[7] X. Liu, M. Li, C. Tang, J. Xia, J. Xiong, L. Liu, M. Klotz, and E. Zhu, 2020.Efficient and effective regularized incomplete multi-view clustering, IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 8, pp. 2634–2646.
[8] M. Hu and S. Chen, 2019.One-pass incomplete multi-view clustering, in Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, pp. 3838–3845.
[9] N. Rai, S. Negi, S. Chaudhury, and O. Deshmukh, 2016.Partial multi-view clustering using graph regularized rmf, in 2016 23rd International Conference on Pattern Recognition ICPR. IEEE, pp. 2192–2197.
[10] Y. Lin, Y. Gou, Z. Liu, B. Li, J. Lv, and X. Peng, 2021.Completer: Incomplete multi-view clustering via contrastive prediction, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11 174–11 183.
[11] L. Tran, X. Liu, J. Zhou, and R. Jun, 2017.Missing modalities imputation via cascaded residual autoencoder, in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1405–1414.
[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, 2017.Unpaired image-to-image translation using cycle-consistent adversarial networks, in Proceedings of the IEEE international conference on computer vision, pp. 2223–2232.
[13] D. Guo, X. Su, Y. Lian, L. Liu, and H. Wang, 2020.Two-stage partial image text clustering tpi-tc, IET Computer Vision.
[14] S. Routray, A. K. Ray, and C. Mishra, 2017.Analysis of various image feature extraction methods against noisy image: Sift, surf and hog, in 2017 Second International Conference on Electrical, Computer and Communication Technologies ICECCT. IEEE, pp. 1–5.
[15] M. Mirza and S. Osindero, 2014.Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784.
[16] M. Hu and S. Chen, 2019.Doubly aligned incomplete multi-view clustering, arXiv preprint arXiv:1903.02785.
[17] H. Zhao, H. Liu, and Y. Fu, 2016.Incomplete multi-modal visual data grouping, in IJCAI, pp. 2392–2398.
[18] F. Bai, A. Trivedi, H. Daume III, and S. L. DuVall, 2010.Multiview clustering with incomplete views, in Proceedings of the NIPS Workshop on Machine Learning for Social Computing. Citeseer.
[19] X. Liu, X. Zhu, M. Li, L. Wang, E. Zhu, T. Liu, M. Klotz, D. Shen, J. Yin, and W. Gao, 2020.Multiple kernel kk-means with incomplete kernels, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 8, pp. 2634–2646.
[20] R. Hadsell, S. Chopra, and Y. LeCun, 2006.Dimensionality reduction by learning an invariant mapping, in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition vol. 2. IEEE, pp. 1735–1742.
[21] S.-Y. Li, Y. Jiang, and Z.-H. Zhou, 2014.Partial multi-view clustering, in Proceedings of the AAAI conference on artificial intelligence, vol. 28, no. 1.
[22] W. Shao, L. He, and P. S. Yu, 2015.Multiple incomplete views clustering via weighted nonnegative matrix factorization with L2,1 regularization, in Joint European conference on machine learning and knowledge discovery in databases. Springer, pp. 318–334.
[24] Q. Wang, Z. Ding, Z. Tao, Q. Gao, and Y. Fu, 2018. Partial multi-view clustering via consistent gan, in 2018 IEEE International Conference on Data Mining ICDM. IEEE, pp. 1290–1295.

[25] Y. W. M. C. B. H. S. Z. Q. H. Pengfei Zhu, Xinjie Yao, 2022. Latent heterogeneous graph network for incomplete multi-view learning. IEEE Transactions on Multimedia, no. 3154592.

[26] Z. Xue, J. Du, C. Zheng, J. Song, W. Ren, and M. Liang, 2021. Clustering-induced adaptive structure enhancing network for incomplete multi-view data. in IJCAI, pp. 3235–3241.

[27] J. Wen, Z. Zhang, Y. Xu, B. Zhang, L. Fei, and G.-S. Xie, 2020. Cdimc-net: Cognitive deep incomplete multi-view clustering network, in Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, C. Bessiere, Ed. International Joint Conferences on Artificial Intelligence Organization, pp. 3230–3236.

[28] C. Buciluța, R. Caruana, and A. Niculescu-Mizil, 2006. Model compression, in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 535–541.

[29] Y. Zhai, Q. Ye, S. Lu, M. Jia, R. Ji, and Y. Tian, 2020. Multiple expert brainstorming for domain adaptive person re-identification, in European Conference on Computer Vision. Springer, pp. 594–611.

[30] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, 2018. Deep mutual learning, in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4320–4328.

[31] X. Liang, L. Wu, J. Li, Y. Wang, Q. Meng, T. Qin, W. Chen, M. Zhang, and T.-Y. Liu, 2021. R-drop: Regularized dropout for neural networks, in Advances in Neural Information Processing Systems. M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc. pp. 10 890–10 905.

[32] V. Nair and G. E. Hinton, 2010. Rectified linear units improve restricted boltzmann machines, in Icml.

[33] L. Van der Maaten and G. Hinton, 2008. Visualizing data using t-sne. Journal of machine learning research, vol. 9, no. 11.