A Two-Stage Information Extraction Network for Incomplete Multi-View Multi-Label Classification

Xin Tan*, Ce Zhao*, Chengliang Liu, Jie Wen†, Zhanyan Tang

Shenzhen Key Laboratory of Visual Object Detection and Recognition, Harbin Institute of Technology, Shenzhen, China
Kevin1205255113@outlook.com, alcor_zhao@outlook.com, liucl1996@163.com, {jiewen_pr, lingf5877}@126.com

Abstract

Recently, multi-view multi-label classification (MvMLC) has received a significant amount of research interest and many methods have been proposed based on the assumptions of view completion and label completion. However, in real-world scenarios, multi-view multi-label data tends to be incomplete due to various uncertainties involved in data collection and manual annotation. As a result, the conventional MvMLC methods fail. In this paper, we propose a new two-stage MvMLC network to solve this incomplete MvMLC issue with partial missing views and missing labels. Different from the existing works, our method attempts to leverage the diverse information from the partially missing data based on the information theory. Specifically, our method aims to minimize task-irrelevant information while maximizing task-relevant information through the principles of information bottleneck theory and mutual information extraction. The first stage of our network involves training view-specific classifiers to concentrate the task-relevant information. Subsequently, in the second stage, the hidden states of these classifiers serve as input for an alignment model, an autoencoder-based mutual information extraction framework, and a weighted fusion classifier to make the final prediction. Extensive experiments performed on five datasets validate that our method outperforms other state-of-the-art methods. Code is available at https://github.com/KevinTan10/TSIEN.

Introduction

An object can be characterized by multiple data from different views, which means these multi-view data may share some instance information while also potentially possessing unique information (Xu et al. 2022; Jiang et al. 2022). Moreover, an object or image may contain rich information, leading to the assignment of multiple labels. For example, a natural landscape image can be described by various feature descriptors such as SIFT, Gist, and HSV, and it may be annotated with multiple labels such as sky, tree, bird, etc. In order to effectively exploit the multi-view multi-label data, multi-view multi-label classification emerges and many methods have been proposed based on the assumption of view completion and label completion in the past years (Zhu et al. 2018; Zhang et al. 2018; Wu et al. 2019). However, in recent years, many reserchers have observed that in real-world applications, the multi-view multi-label data are often incomplete due to various uncertainties involved in data collection and manual annotation (Li, Wan, and He 2021; Xu et al. 2021; Wang et al. 2021). The absence of views and labels has a significant impact on MvMLC (Xu, Tao, and Xu 2015; Liu et al. 2021). Consequently, the investigation of incomplete multi-view multi-label classification (iMvMLC) is imperative.

For iMvMLC, traditional methods such as iMvWL (Tan et al. 2018) and NAIML (Li and Chen 2021) have been proposed. iMvWL performs simultaneous mapping of multi-view features and multi-label information into a discriminative shared subspace. NAIML effectively leverages the consensus of multiple views and the structures of multiple labels. Despite achieving certain outcomes, these methods have some drawbacks. For example, the two traditional methods are all shallow machine learning methods, which cannot capture the underlying discriminative information of data and are all computationally inefficient. Moreover, the methods based on matrix factorization like iMvWL cannot handle the new-coming test samples with the trained model. More recently, methods based on deep neural networks (DNN), such as LMVCAT (Liu et al. 2023b) and DICNet (Liu et al. 2023a), have demonstrated remarkable advantages in this task. Compared to traditional methods such as matrix factorization, DNN are better at capturing high-level semantic information and are more suitable for complex iMvMLC tasks (Wen et al. 2020; Liu et al. 2023a).

Focusing on DNN architectures for iMvMLC, in addition to the widely adopted autoencoder, contrastive learning is a promising approach to improve the performance (Liu et al. 2023a). However, existing iMvMLC methods have not delved into the essence of contrastive learning, specifically increasing cross-view mutual information (MI) to extract task-relevant information. This limitation also results in their models not being able to fully extract useful information. To overcome this bottleneck, in this paper, we propose an information extraction network for incomplete multi-view multi-label classification from an information-theoretic perspective. Different from the existing works, the proposed model seeks to integrate the information bottleneck (IB) and mutual information extraction to address the iMvMLC prob-

*These authors contributed equally.
†Corresponding author.

lem. Our main contributions are outlined as follows:

- This is the first deep neural network architecture designed entirely from an information-theoretic perspective to address the iMvMLC problem. Our work demonstrates the promising prospects of information theory for solving the iMvMLC problem.
- In this paper, we propose a new information theory based framework to balance the trade-off of cross-view MI and view-specific information. Very different from the existing works, the proposed method can extract the task-relevant information from each single view and at the same time sufficiently consider the shared and complementary information across multiple views in a joint framework.

Preliminaries

In this section, we mainly introduce the definition of iMvMLC and some notations used throuth the paper. In addition, some related works about two basic information theories used in our method are analyzed.

Problem Definition and Notation Introduction

Definition of iMvMLC A given dataset includes m views and n samples can be denoted as: $\{\mathrm{V}^{(k)} \in \mathbb{R}^{n\times d_{\mathrm{v}}^{(k)}}\}_{k=1}^{m}$, where $d_{\mathrm{v}}^{(k)}$ is the feature dimension of the k-th view. And we define $\mathrm{Y} \in \{0,1\}^{n\times l}$ as the label matrix, where l is the number of categories. $\mathrm{Y}_{i,j} = 1$ indicates that the i-th sample is marked with the j-th category, otherwise $\mathrm{Y}_{i,j} = 0$. Considering the compatibility for missing views and missing labels, we define two key prior matrices, the missing-view indicator $\mathrm{M} \in \{0,1\}^{n\times m}$ and the missing-label indicator $\mathrm{G} \in \{0,1\}^{n\times l}$. Taking the i-th sample as an example, $\mathrm{M}_{i,j} = 1$ means its j-th view is available, otherwise $\mathrm{M}_{i,j} = 0$. $\mathrm{G}_{i,j} = 0$ indicates that we are not sure whether the i-th sample has the j-th label, otherwise $\mathrm{G}_{i,j} = 1$. For simplicity, we fill random values for the missing views in the original data and set '0' for unknown labels. **The task of iMvMLC is** to train a discriminative model on such partially labeled incomplete multi-view data to perform inference on unlabeled test samples with complete or incomplete views.

Representative notations Some representative notations are given in Table 1. Please note that $\mathrm{v}^{(k)}$, $\mathrm{x}^{(k)}$, $\mathrm{z}^{(k)}$, and y represent vectors, while $\mathrm{V}^{(k)}$, $\mathrm{X}^{(k)}$, $\mathrm{Z}^{(k)}$, and Y represent data matrices. $p(\mathrm{x}^{(k)})$ and $q(\mathrm{x}^{(k)})$ are used to denote the real distribution and the approximate distribution of $\mathrm{x}^{(k)}$, respectively. $p(\mathrm{x}^{(u)}|\mathrm{z}^{(k)})$ is a conditional distribution.

Related Works

Information Bottleneck Theory The information bottleneck (IB) theory (Tishby, Pereira, and Bialek 2000) is an appealing method for representation learning. It provides a formula to balance the trade-off between prediction accuracy and information capacity to obtain the representation with reduced task-irrelevant information. To apply the IB theory to DNN, different computationally feasible bounds of MI have been derived (Alemi et al. 2017; Belghazi et al. 2018). However, these works are all specific to a single view. Recently, some works have extended the IB theory into a multi-view form (Wang et al. 2019; Zhang et al. 2022). One study has also tackled the challenge of missing views (Lee and Van der Schaar 2021). However, these methods do not fully consider the shared and complementary information of the multi-view data.

n, m, l	number of samples, views, and categories, respectively
$d_{\mathrm{v}}^{(k)}, d_{\mathrm{x}}^{(k)}, d_{\mathrm{z}}^{(k)}$	different dimensions
$\mathrm{v}^{(k)} \in \mathbb{R}^{d_{\mathrm{v}}^{(k)}}$	the original instance from k-th view, r.v.
$\mathrm{x}^{(k)} \in \mathbb{R}^{d_{\mathrm{x}}^{(k)}}$	first-stage representation, r.v.
$\mathrm{z}^{(k)} \in \mathbb{R}^{d_{\mathrm{z}}^{(k)}}$	second-stage representation, r.v.
$\mathrm{y} \in \{0,1\}^{l}$	label, r.v.
$\mathrm{V}^{(k)}, \mathrm{X}^{(k)}, \mathrm{Z}^{(k)}$	data matrices
$\hat{\mathrm{X}}^{(u,k)} \in \mathbb{R}^{n\times d_{\mathrm{x}}^{(u)}}$	reconstruction of $\mathrm{X}^{(u)}$ from $\mathrm{Z}^{(k)}$
$\mathrm{Y} \in \{0,1\}^{n\times l}$	label matrix
$\hat{\mathrm{Y}}, \hat{\mathrm{Y}}^{*} \in [0,1]^{n\times l}$	prediction of Y
$\mathrm{M} \in \{0,1\}^{n\times m}$	missing-view indicator matrix
$\mathrm{G} \in \{0,1\}^{n\times l}$	missing-label indicator matrix
Φ_E, Φ_D	encoder and decoder, respectively
Ψ_C, Ψ_C^{*}	classifier

Table 1: Notations

Information Theory in Contrastive Learning For contrastive learning, numerous researchers have endeavored to analyze its mechanisms (Wu et al. 2018; Oord, Li, and Vinyals 2018; Tian et al. 2020). Oord et al. found that the Noise-Contrastive Estimation (NCE) loss (Gutmann and Hyvärinen 2010) optimized in contrastive learning is a lower bound of cross-view MI, so they call it InfoNCE loss (Oord, Li, and Vinyals 2018). This provides an insight that the effectiveness of contrastive learning may be attributed to the increase in cross-view MI. Subsequently, by leveraging information theory, a rigorous proof was provided (Tsai et al. 2020). In the context of the multi-view assumption (Sridharan and Kakade 2008), the non-shared information between views is considered approximately redundant, therefore increasing cross-view MI implies the extraction of task-relevant information. However, not all domains' data conform to this assumption, such as the multi-view data, as the data is not obtained through data augmentation. In this case, it is necessary to increase the MI between the representation and the input (Wang et al. 2022). This indicates that the optimal representation lies in finding a balance between increasing cross-view MI and preserving the original information.

Methodology

Inspired by the analysis of information bottleneck theory and constrastive learning, in this paper, we propose a new information theory driven incomplete multi-view multi-label classification network, whose main structure is shown in Figure 1. Taking into account the incomplete nature of labels and views, and considering the diverse characteristics

of multi-views, during training, we not only prioritize classification performance but also emphasize the importance of representation quality. To this end, we design a two-stage network for the iMvMLC task: enhanced representations generation network and incomplete multi-view classification network with mutual prediction.

First Stage: Enhanced Representations Generation Network

Many works have shown that a better representation should contain more task-relevant information while less task-irrelevant information, *i.e.*, the minimal sufficient statistic (Soatto and Chiuso 2014). Generally speaking, in multi-view multi-label classification task, the raw data often has much task-irrelevant (redundant) information, which will degrade the performance of the main classification network.

To address this issue, inspired by (Tishby, Pereira, and Bialek 2000), we propose to learn a better task-relevant representation $\mathrm{x}^{(k)}$ that satisfies the following IB principle to replace the low purity instance $\mathrm{v}^{(k)}$:

$$\min_{\mathrm{x}^{(k)}} -I(\mathrm{x}^{(k)};\mathrm{y}) + \beta^{(k)} I(\mathrm{x}^{(k)};\mathrm{v}^{(k)}) \tag{1}$$

where $\beta^{(k)} > 0$ is a hyper-parameter. $I(\mathrm{x}^{(k)};\mathrm{y})$ indicates the MI between $\mathrm{x}^{(k)}$ and label y, which can be approximated as reducing the common cross-entropy loss (Achille and Soatto 2018; Amjad and Geiger 2019). $I(\mathrm{x}^{(k)};\mathrm{v}^{(k)})$ is the MI between $\mathrm{x}^{(k)}$ and instance $\mathrm{v}^{(k)}$, which can be regarded as a constraint that ensures $\mathrm{x}^{(k)}$ to contain less task-irrelevant information. However, it appears to pose computational intractability. In our research, we employ a variational approximation technique to mitigate this challenge, as outlined in the work of Alemi et al. (Alemi et al. 2017). Concretely, we deduce the subsequent upper bound:

$$I(\mathrm{x}^{(k)};\mathrm{v}^{(k)}) \leq \mathbb{E}_{p(\mathrm{x}^{(k)},\mathrm{v}^{(k)})}[\log \frac{p(\mathrm{x}^{(k)}|\mathrm{v}^{(k)})}{q(\mathrm{x}^{(k)})}] \tag{2}$$

where $\mathbb{E}$ represents expectation, and $q(\mathrm{x}^{(k)})$ represents an approximate distribution to marginal distribution $p(\mathrm{x}^{(k)})$. Focusing on this upper bound, we can arbitrarily set $q(\mathrm{x}^{(k)}) = \mathcal{N}(\mathrm{x}^{(k)}|\mathbf{0},\mathbf{I})$ and $p(\mathrm{x}^{(k)}|\mathrm{v}^{(k)}) = \mathcal{N}(\mathrm{x}^{(k)}|\mu^{(k)},\Sigma^{(k)})$ for computational expedience, where both mean $\mu^{(k)} \in \mathbb{R}^{d_{\mathrm{x}}^{(k)}}$ and diagonal covariance $\Sigma^{(k)} \in \mathbb{R}^{d_{\mathrm{x}}^{(k)} \times d_{\mathrm{x}}^{(k)}}$ are correlated with $\mathrm{v}^{(k)}$. Specifically, we utilize a special MLP $\Phi_F^{(k)}$ to derive $\mu^{(k)}$ and $\Sigma^{(k)}$. When the input of $\Phi_F^{(k)}$ is the data matrix $\mathrm{V}^{(k)} \in \mathbb{R}^{n \times d_{\mathrm{v}}^{(k)}}$, let we use two new symbols to represent its output, namely $\mathrm{U}^{(k)} \in \mathbb{R}^{n \times d_{\mathrm{x}}^{(k)}}$ and $\mathrm{S}^{(k)} \in \mathbb{R}^{n \times d_{\mathrm{x}}^{(k)}}$, respectively. So $\mathrm{X}^{(k)} = \mathrm{U}^{(k)} + \epsilon \odot \mathrm{S}^{(k)}$ (Kingma and Welling 2014), where $\epsilon \in \mathbb{R}^{n \times d_{\mathrm{x}}^{(k)}}$ is a standard Gaussian distribution matrix, and $\odot$ denotes the Hadamard product. By substituting the two Gaussians $p(\mathrm{x}^{(k)}|\mathrm{v}^{(k)})$ and $q(\mathrm{x}^{(k)})$ into formula (2), decomposing them into multiple one-dimensional Gaussians, and approximating the marginal distribution $p(\mathrm{v}^{(k)})$ using an empirical distribution, formula (2) can be transformed into the equivalent problem $\min \frac{1}{2nd_{\mathrm{x}}^{(k)}} \sum_{i=1}^{n} \sum_{j=1}^{d_{\mathrm{x}}^{(k)}} (-\log \mathrm{S}_{i,j}^{(k)^2} + \mathrm{U}_{i,j}^{(k)^2} + \mathrm{S}_{i,j}^{(k)^2} - 1)$.

Eventually, with incorporating missing indicator matrices M and G to exclude the negative influences of the unavailable views and labels to the model training, the objective function of the network in the first stage becomes:

$$\begin{aligned} & L_{IB}^{(k)} \\ & = \frac{1}{nl} \sum_{i=1}^{n} \sum_{j=1}^{l} \mathrm{M}_{i,k} \mathrm{G}_{i,j} [(1 - \mathrm{Y}_{i,j}) \log(1 - \hat{\mathrm{Y}}_{i,j}^{(k)}) + \\ & \quad \mathrm{Y}_{i,j} \log(\hat{\mathrm{Y}}_{i,j}^{(k)})] + \frac{\beta^{(k)}}{2nd_{\mathrm{x}}^{(k)}} \sum_{i=1}^{n} \sum_{j=1}^{d_{\mathrm{x}}^{(k)}} \mathrm{M}_{i,k} (-\log \mathrm{S}_{i,j}^{(k)^2} + \\ & \quad \mathrm{U}_{i,j}^{(k)^2} + \mathrm{S}_{i,j}^{(k)^2} - 1) \end{aligned} \tag{3}$$

where $\hat{\mathrm{Y}}^{(k)}$ is the output of module $\Psi_C^{(k)}$. The first term is a cross-entropy term allowing for the extraction of task-relevant information, corresponding to $-I(\mathrm{x}^{(k)};\mathrm{y})$. The second term can be regarded as an information term aimed at compressing the information, corresponding to $I(\mathrm{x}^{(k)};\mathrm{v}^{(k)})$.

Clearly, the training in the first stage actually involves separately utilizing data from each view to predict the label. As shown in Figure 1, the raw data $\mathrm{V}^{(k)}$ directly pass through the two MLPs $\Phi_F^{(k)}$ and $\Psi_C^{(k)}$ to predict the label Y in the first stage. $\Psi_C^{(k)}$ is used solely for training $\Phi_F^{(k)}$ and will be discarded after training. After the first-stage training, we obtain several MLPs $\{\Phi_F^{(k)}\}_{k=1}^{m}$ that can enhance the concentration of task-relevant information.

Second Stage: Incomplete Multi-view Classification Network with Mutual Prediction

Mutual Prediction To further exploit the useful information, one way is contrastive learning that optimizes an original InfoNCE loss (Liu et al. 2023a), but it faces the risk of pushing the positive samples away. Supervised contrastive learning (Khosla et al. 2020) is a superior approach, yet is hard to be adopted in iMvMLC due to the incomplete label. Inspired by (Tsai et al. 2020; Wang et al. 2022), we propose an autoencoder-based mutual information extraction framework, which can perform mutual prediction between views.

Inspired by former works, for multi-view learning, we seek to increase the cross-view MI while preserving some of the original information based on the information theory as follows:

$$\begin{aligned} \max_{\mathrm{z}^{(k)},k=1,\ldots m} & \frac{1}{m(m-1)} \sum_{k=1}^{m} \sum_{u \neq k} I(\mathrm{z}^{(k)};\mathrm{x}^{(u)}) + \\ & \frac{\lambda}{m} \sum_{k=1}^{m} I(\mathrm{z}^{(k)};\mathrm{x}^{(k)}) \end{aligned} \tag{4}$$

where $\lambda > 0$ is a combination factor, $\mathrm{z}^{(k)}$ is the second-stage representation. According to the definition of MI (Shannon

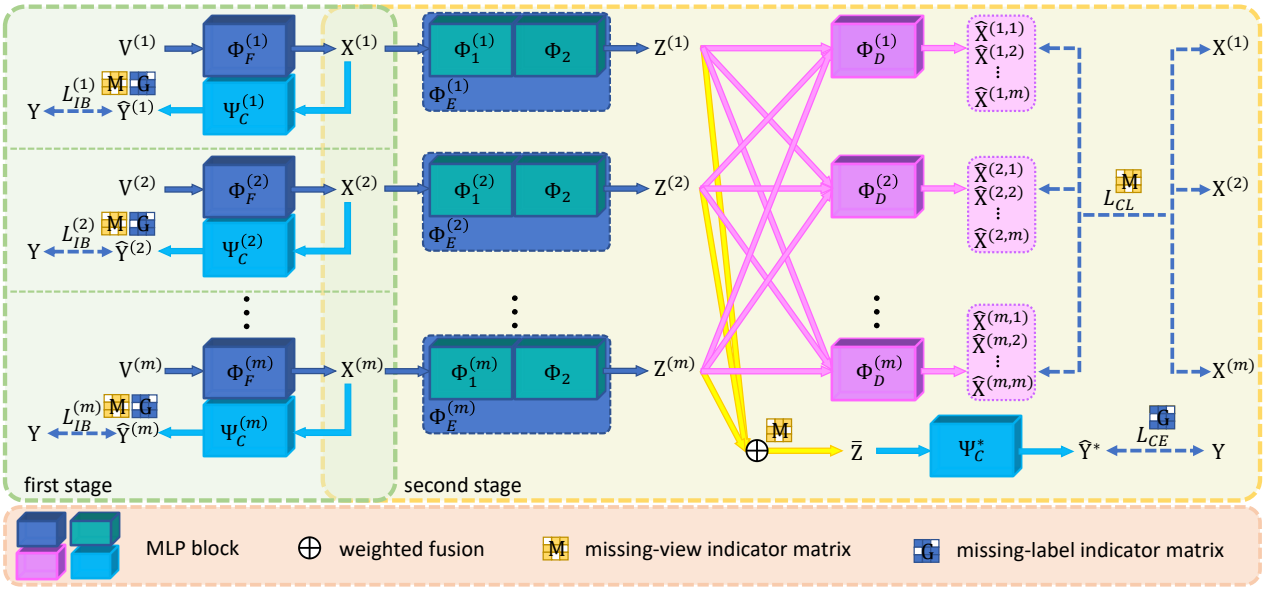


Figure 1: An overview of our two-stage network.

1948), we have $I(\mathrm{z}^{(k)};\mathrm{x}^{(u)}) = H(\mathrm{x}^{(u)}) - H(\mathrm{x}^{(u)}|\mathrm{z}^{(k)})$, where $H(\mathrm{x}^{(u)})$ represents the entropy and $H(\mathrm{x}^{(u)}|\mathrm{z}^{(k)})$ represents the conditional entropy. Since the m trained MLPs $\{\Phi_F^{(k)}\}_{k=1}^m$ are fixed in the second stage, we can ignore $H(\mathrm{x}^{(u)})$. So our goal is to maximize $-H(\mathrm{x}^{(u)}|\mathrm{z}^{(k)})$. By employing variational approximation, we have

$$-H(\mathrm{x}^{(u)}|\mathrm{z}^{(k)}) \geq \mathbb{E}_{p(\mathrm{x}^{(u)},\mathrm{z}^{(k)})}[\log q(\mathrm{x}^{(u)}|\mathrm{z}^{(k)})] \tag{5}$$

Similarly, we can also set $q(\mathrm{x}^{(u)}|\mathrm{z}^{(k)}) = \mathcal{N}(\mathrm{x}^{(u)}|\mu^{(u,k)}, \sigma\mathbf{I})$ with a fixed $\sigma \in \mathbb{R}$, and the mean value $\mu^{(u,k)} \in \mathbb{R}^{d_{\mathrm{x}}^{(u)}}$ is dependent on $\mathrm{z}^{(k)}$. In this case, maximizing $\mathbb{E}_{p(\mathrm{x}^{(u)},\mathrm{z}^{(k)})}[\log q(\mathrm{x}^{(u)}|\mathrm{z}^{(k)})]$ is equivalent to minimizing $\mathbb{E}_{p(\mathrm{x}^{(u)},\mathrm{z}^{(k)})}[\|\mathrm{x}^{(u)} - \mu^{(u,k)}\|_2^2]$. In our work, we employ a decoder $\Phi_D^{(u)}$ to generate $\mu^{(u,k)}$, *i.e.*, $\mu^{(u,k)} = \Phi_D^{(u)}(\mathrm{z}^{(k)})$ ($\hat{\mathrm{X}}^{(u,k)} = \Phi_D^{(u)}(\mathrm{Z}^{(k)})$). With the same treatment applied to $I(\mathrm{z}^{(k)};\mathrm{x}^{(k)})$, by introducing M, we can obtain the following information theory based loss for incomplete multi-view learning:

$$\begin{aligned} L_{CL} &= \frac{1}{nm(m-1)} \sum_{i=1}^{n} \sum_{k=1}^{m} \sum_{u \neq k} \frac{\mathrm{M}_{i,u}\mathrm{M}_{i,k}}{d_{\mathrm{x}}^{(u)}} \left\| \mathrm{X}_{i,\cdot}^{(u)} - \hat{\mathrm{X}}_{i,\cdot}^{(u,k)} \right\|_2^2 \\ &\quad + \frac{\lambda}{nm} \sum_{i=1}^{n} \sum_{k=1}^{m} \frac{\mathrm{M}_{i,k}}{d_{\mathrm{x}}^{(k)}} \left\| \mathrm{X}_{i,\cdot}^{(k)} - \hat{\mathrm{X}}_{i,\cdot}^{(k,k)} \right\|_2^2 \\ &= L_{CL1} + \lambda L_{CL2} \end{aligned} \tag{6}$$

where L_{CL1} represents the first term, which is used to increase the cross-view MI. λL_{CL2} corresponds to the second term, representing the preservation of the original information. Note that minimizing loss (6) can further enhance the cross-view MI in the network of the second stage similar to the contrastive learning, but it adopts a non-instance discrimination perspective different from InfoNCE loss.

Now we are discussing the structure of the new autoencoder network in this stage. Each encoder $\Phi_E^{(k)}$ comprises two parts. The first part is a view-specific module $\Phi_1^{(k)}$ mapping $\mathrm{X}^{(k)}$ to a joint semantic space and the second part is a shared module Φ_2. Therefore $\Phi_E^{(k)}(\mathrm{X}^{(k)}) = \Phi_2(\Phi_1^{(k)}(\mathrm{X}^{(k)}))$, where $\Phi_1^{(k)}$ and Φ_2 are both MLP. Furthermore, we utilize one specific $\Phi_D^{(u)}$ in conjunction with different representations $\{\mathrm{Z}^{(k)}\}_{k=1}^m$ to reconstruct specific $\mathrm{X}^{(u)}$, so that $\hat{\mathrm{X}}^{(u,k)} = \Phi_D^{(u)}(\mathrm{Z}^{(k)})$. In this way, we use the same set of decoders and part of the same encoder Φ_2 to implicitly align the representations.

One thing that needs to be pointed out is that we reduce the MI between $\mathrm{x}^{(k)}$ and $\mathrm{v}^{(k)}$ in the first stage while increasing the MI between $\mathrm{z}^{(k)}$ and $\mathrm{x}^{(k)}$ in the second stage. This may appear contradictory, but we provide a valid justification why the method works. When utilizing autoencoder, $\mathrm{z}^{(k)}$ strives to preserve as much information from $\mathrm{x}^{(k)}$ as possible, making the efficiency of $\mathrm{x}^{(k)}$ highly significant. So, for high-noise/redundant original instance $\mathrm{v}^{(k)}$, it is necessary to perform an initial step of information concentration especially the target task-relevant information extraction before preserving the original information, then perform the reconstruction of $\mathrm{x}^{(k)}$ rather than $\mathrm{v}^{(k)}$. In practice, this is contingent upon the quality of $\mathrm{v}^{(k)}$.

Final Prediction and Overall Objective Loss Now that we have obtained the informative representations $\{\mathrm{Z}^{(k)}\}_{k=1}^m$ from different views, to make the final prediction, we need to fuse these representations. While concatenation is unavailable for the incompleteness, one simple method is to aver-

age, but it does not take into account the varying importance of each view. Hence, we adopt a commonly used method for handling missing views in multi-view fusion as (Wen et al. 2020; Trosten et al. 2021):

$$\bar{\mathrm{Z}}_{i,\cdot} = \sum_{k=1}^{m} w^{(k)} \frac{\mathrm{Z}_{i,\cdot}^{(k)} \mathrm{M}_{i,k}}{\sum_{r=1}^{m} \mathrm{M}_{i,r}} \tag{7}$$

which is a weighted average of the $\mathrm{Z}_{i,\cdot}^{(k)}$. $\{w^{(k)}\}_{k=1}^{m}$ are learnable weights that have been processed through softmax.

For the final classifier Ψ_C^*, we employ MLPs. And we utilize the cross-entropy loss function, which is showed as follows:

$$\begin{aligned} L_{CE} = \frac{1}{nl} \sum_{i=1}^{n} \sum_{j=1}^{l} \mathrm{G}_{i,j} [(1 - \mathrm{Y}_{i,j}) \log(1 - \hat{\mathrm{Y}}_{i,j}^*) + \\ \mathrm{Y}_{i,j} \log(\hat{\mathrm{Y}}_{i,j}^*)] \end{aligned} \tag{8}$$

where the final prediction $\hat{\mathrm{Y}}^* = \Psi_C^*(\bar{\mathrm{Z}}) \in [0,1]^{n \times l}$.

By combining (6) and (8) through a hyper-parameter α, we can obtain the final objective loss for the network in the second stage:

$$\begin{aligned} L_{SCL} &= L_{CE} + \alpha(L_{CL1} + \lambda L_{CL2}) \\ &= L_{CE} + \alpha L_{CL1} + \gamma L_{CL2} \end{aligned} \tag{9}$$

Algorithm 1: Training Process

Input: Incomplete multi-view data $\{\mathrm{V}^{(k)}\}_{k=1}^{m}$ with missing-view indicator matrix M, and corresponding multi-label matrix Y with missing-label indicator matrix G; batch size *B*; Hyper-parameters $\{\beta^{(k)}\}_{k=1}^{m}$, α and γ
Output: The trained model

procedure STAGE 1
Initialize MLPs $\{\Phi_F^{(k)}\}_{k=1}^{m}$, classifiers $\{\Psi_C^{(k)}\}_{k=1}^{m}$;
for k=1 **to** m **do**
  while *not converged* **do**
    compute $\mathrm{X}^{(k)} = \Phi_F^{(k)}(\mathrm{V}^{(k)})$, $\hat{\mathrm{Y}} = \Psi_C^{(k)}(\mathrm{X}^{(k)})$, $L_{IB}^{(k)}$;
    update $\Phi_F^{(k)}$, $\Psi_C^{(k)}$;
  end while
end for
Save $\{\Phi_F^{(k)}\}_{k=1}^{m}$;
end procedure
procedure STAGE 2
Initialize MLPs $\{\Phi_1^{(k)}\}_{k=1}^{m}$, Φ_2, decoders $\{\Phi_D^{(k)}\}_{k=1}^{m}$, classifier Ψ_C^* and $\{w^{(k)}\}_{k=1}^{m}$.
while *not converged* **do**
  for k=1 **to** *m* **do**
    compute $\mathrm{Z}^{(k)} = \Phi_2(\Phi_1^{(k)}(\Phi_F^{(k)}(\mathrm{V}^{(k)})))$;
  end for
  for u=1 **to** *m* **do**
    for k=1 **to** *m* **do**
      compute $\hat{\mathrm{X}}^{(u,k)} = \Phi_D^{(u)}(\mathrm{Z}^{(k)})$;
    end for
  end for
  compute $\bar{\mathrm{Z}}$, $\hat{\mathrm{Y}}^* = \Psi_C^*(\bar{\mathrm{Z}})$, and L_{SCL};
  update $\{\Phi_1^{(k)}, \Phi_D^{(k)}, w^{(k)}\}_{k=1}^{m}$, Ψ_C^*, and Φ_2;
end while
Save Φ_2, $\{\Phi_1^{(k)}\}_{k=1}^{m}$, Ψ_C^*, and $\{w^{(k)}\}_{k=1}^{m}$;
end procedure

Experiments

Datasets and Evaluation Metrics

Following the datasets selection by (Tan et al. 2018; Li and Chen 2021; Wen et al. 2023), we conduct experiments on five commonly used datasets, namely Corel5k (Duygulu et al. 2002), Pascal07 (Everingham et al. 2010), ESPGame (Von Ahn and Dabbish 2004), IAPRTC12 (Grubinger et al. 2006), and Mirflickr (Huiskes and Lew 2008). These datasets all encompass six distinct views or perspectives, including GIST, HSV, Hue, Sift, RGB, and LAB. To simulate the real-world scenario of missing views, we randomly remove 50% of the data in each view under the condition that each sample contains at least one available view. Then we randomly select 70% of samples as training data. In addition, for the training data, 50% of positive tags and 50% of negative tags within each class are randomly masked/deleted.

We evaluate the models using six multi-label classification related metrics: Average Precision (AP), Hamming Loss (HL), Ranking Loss (RL), adapted area under curve (AUC), OneError (OE), and Coverage (Cov) in our experiments. For more details, one can refer to (Zhang and Zhou 2013). For convenience, $1-$HL and $1-$RL are used as substitutes for HL and RL, respectively, so that higher values indicate better performance of the model.

Method Comparison and Implementation Details

We compare our proposed method with six competitive methods in the field. The first category involves traditional non-DNN iMvMLC models iMvWL (Tan et al. 2018) and NAIML (Li and Chen 2021), which have been introduced in the introduction section. The second category consists of DNN-based models CDMM (Zhao et al. 2021) and DeepIMV (Lee and Van der Schaar 2021). CDMM employs DNN to solve the problem of consistency and diversity among views. The latter method DeepIMV employs IB framework to derive marginal and joint representations from the given data with the product-of-experts strategy. However, these two methods are unable to handle missing views or mutliple labels. Hence, we will apply mean imputation or appropriately modify them. The final category includes DNN-based iMvMLC models LMVCAT (Liu et al. 2023b) and DICNet (Liu et al. 2023a). Among them, LMVCAT is a transformer-based method; DICNet is a simple contrastive learning framework with InfoNCE loss.

For the aforementioned methods, all settings adhere to their optimal configurations or undergo only necessary adjustments. For our proposed method, one notable drawback is that as the number of views increases, adjusting the hyper-parameters $\{\beta^{(k)}\}_{k=1}^{m}$ becomes more challenging. In our experiments, $\{\beta^{(k)}\}_{k=1}^{m}$ will be set to be equal, *i.e.*, $\beta^{(1)} = \beta^{(2)} = \cdots = \beta^{(m)} = \beta$. Similar to other DNN-based meth-

DATASET	METRIC	iMvWL	NAIML	CDMM	DeepIMV	LMVCAT	DICNet	OURS
Corel5k	AP ↑	$0.283_{0.008}$	$0.309_{0.004}$	$0.354_{0.004}$	$0.376_{0.011}$	$0.382_{0.004}$	$0.381_{0.004}$	$\mathbf{0.436}_{0.007}$
	1−HL ↑	$0.978_{0.000}$	$0.987_{0.000}$	$0.987_{0.000}$	$0.987_{0.000}$	$0.986_{0.000}$	$\mathbf{0.988}_{0.000}$	$\mathbf{0.988}_{0.000}$
	1−RL ↑	$0.865_{0.005}$	$0.878_{0.002}$	$0.884_{0.003}$	$0.863_{0.005}$	$0.880_{0.002}$	$0.882_{0.004}$	$\mathbf{0.917}_{0.002}$
	AUC ↑	$0.868_{0.005}$	$0.881_{0.002}$	$0.888_{0.003}$	$0.866_{0.005}$	$0.883_{0.002}$	$0.884_{0.004}$	$\mathbf{0.920}_{0.002}$
	OE ↓	$0.689_{0.015}$	$0.650_{0.009}$	$0.590_{0.007}$	$0.539_{0.015}$	$0.547_{0.006}$	$0.532_{0.007}$	$\mathbf{0.487}_{0.015}$
	Cov ↓	$0.298_{0.008}$	$0.275_{0.005}$	$0.277_{0.007}$	$0.298_{0.010}$	$0.273_{0.006}$	$0.273_{0.011}$	$\mathbf{0.194}_{0.006}$
Pascal07	AP ↑	$0.437_{0.018}$	$0.488_{0.003}$	$0.508_{0.005}$	$0.548_{0.008}$	$0.519_{0.006}$	$0.505_{0.012}$	$\mathbf{0.581}_{0.009}$
	1−HL ↑	$0.882_{0.004}$	$0.928_{0.001}$	$0.931_{0.001}$	$0.930_{0.001}$	$0.924_{0.003}$	$0.929_{0.001}$	$\mathbf{0.934}_{0.001}$
	1−RL ↑	$0.736_{0.015}$	$0.783_{0.001}$	$0.812_{0.004}$	$0.815_{0.008}$	$0.811_{0.004}$	$0.783_{0.008}$	$\mathbf{0.849}_{0.005}$
	AUC ↑	$0.767_{0.015}$	$0.811_{0.001}$	$0.838_{0.003}$	$0.835_{0.009}$	$0.834_{0.004}$	$0.809_{0.006}$	$\mathbf{0.868}_{0.004}$
	OE ↓	$0.638_{0.023}$	$0.579_{0.006}$	$0.581_{0.008}$	$0.537_{0.014}$	$0.579_{0.006}$	$0.573_{0.015}$	$\mathbf{0.509}_{0.011}$
	Cov ↓	$0.323_{0.015}$	$0.273_{0.002}$	$0.241_{0.003}$	$0.232_{0.009}$	$0.237_{0.005}$	$0.269_{0.006}$	$\mathbf{0.197}_{0.005}$
ESPGame	AP ↑	$0.244_{0.005}$	$0.246_{0.002}$	$0.289_{0.003}$	$0.294_{0.004}$	$0.294_{0.004}$	$0.297_{0.002}$	$\mathbf{0.319}_{0.004}$
	1−HL ↑	$0.972_{0.000}$	$0.983_{0.000}$	$0.983_{0.000}$	$0.982_{0.000}$	$0.982_{0.000}$	$\mathbf{0.983}_{0.000}$	$\mathbf{0.983}_{0.000}$
	1−RL ↑	$0.808_{0.002}$	$0.818_{0.002}$	$0.832_{0.001}$	$0.832_{0.002}$	$0.828_{0.002}$	$0.832_{0.001}$	$\mathbf{0.859}_{0.002}$
	AUC ↑	$0.813_{0.002}$	$0.824_{0.002}$	$0.836_{0.001}$	$0.835_{0.002}$	$0.833_{0.002}$	$0.836_{0.001}$	$\mathbf{0.863}_{0.002}$
	OE ↓	$0.657_{0.013}$	$0.661_{0.003}$	$0.604_{0.005}$	$0.567_{0.008}$	$0.566_{0.009}$	$0.561_{0.007}$	$\mathbf{0.546}_{0.007}$
	Cov ↓	$0.452_{0.004}$	$0.429_{0.003}$	$0.426_{0.004}$	$0.394_{0.004}$	$0.410_{0.004}$	$0.407_{0.003}$	$\mathbf{0.349}_{0.003}$
IAPRTC12	AP ↑	$0.237_{0.003}$	$0.261_{0.001}$	$0.305_{0.004}$	$0.325_{0.004}$	$0.317_{0.003}$	$0.323_{0.001}$	$\mathbf{0.361}_{0.004}$
	1−HL ↑	$0.969_{0.000}$	$0.980_{0.000}$	$0.981_{0.000}$	$0.980_{0.000}$	$0.980_{0.000}$	$\mathbf{0.981}_{0.000}$	$\mathbf{0.981}_{0.000}$
	1−RL ↑	$0.833_{0.002}$	$0.848_{0.001}$	$0.862_{0.002}$	$0.873_{0.004}$	$0.870_{0.001}$	$0.873_{0.001}$	$\mathbf{0.898}_{0.003}$
	AUC ↑	$0.835_{0.001}$	$0.850_{0.001}$	$0.864_{0.002}$	$0.875_{0.004}$	$0.872_{0.001}$	$0.874_{0.000}$	$\mathbf{0.899}_{0.002}$
	OE ↓	$0.648_{0.008}$	$0.610_{0.005}$	$0.568_{0.008}$	$0.543_{0.008}$	$0.557_{0.005}$	$0.532_{0.002}$	$\mathbf{0.505}_{0.007}$
	Cov ↓	$0.436_{0.005}$	$0.408_{0.004}$	$0.403_{0.004}$	$0.335_{0.007}$	$0.352_{0.003}$	$0.351_{0.001}$	$\mathbf{0.291}_{0.006}$
Mirflickr	AP ↑	$0.490_{0.012}$	$0.551_{0.002}$	$0.570_{0.002}$	$0.612_{0.005}$	$0.594_{0.005}$	$0.589_{0.005}$	$\mathbf{0.631}_{0.003}$
	1−HL ↑	$0.839_{0.002}$	$0.882_{0.001}$	$0.886_{0.001}$	$0.887_{0.001}$	$0.882_{0.002}$	$0.888_{0.002}$	$\mathbf{0.895}_{0.001}$
	1−RL ↑	$0.803_{0.008}$	$0.844_{0.001}$	$0.856_{0.001}$	$0.871_{0.002}$	$0.865_{0.003}$	$0.863_{0.004}$	$\mathbf{0.887}_{0.001}$
	AUC ↑	$0.787_{0.012}$	$0.837_{0.001}$	$0.846_{0.001}$	$0.856_{0.003}$	$0.853_{0.003}$	$0.849_{0.004}$	$\mathbf{0.872}_{0.001}$
	OE ↓	$0.489_{0.022}$	$0.415_{0.003}$	$0.369_{0.004}$	$0.331_{0.007}$	$0.358_{0.008}$	$0.363_{0.007}$	$\mathbf{0.321}_{0.006}$
	Cov ↓	$0.428_{0.013}$	$0.369_{0.002}$	$0.360_{0.001}$	$0.323_{0.003}$	$0.333_{0.003}$	$0.348_{0.007}$	$\mathbf{0.305}_{0.003}$

Table 2: The performance of different methods on various datasets. The best results are highlighted in bold.

ods (Lee and Van der Schaar 2021; Liu et al. 2023b), we introduce dropout layers in our MLPs. Furthermore, the final layer of $\Phi_1^{(k)}$ and Φ_2 is augmented with batch normalization. Our training process is illustrated in Algorithm 1.

Experimental Results and Analysis

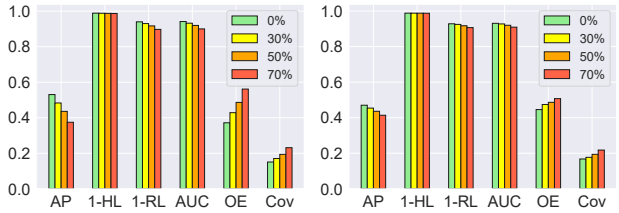
Table 2 presents the performance of each method on the six datasets, indicating the mean and variance for different metrics. Part of the experimental results are from (Liu et al. 2023c). All experiments were conducted with 10 repetitions to ensure accuracy and statistical significance. From the table, it is evident that:

- Compared to other methods, our proposed information theory-based method demonstrates superior classification performance across all datasets and ranks first in all metrics. This indicates that information theory is an effective approach for the iMvMLC task. In terms of the most representative metric, AP, our method exhibits improvements ranging from 2% to 5%. Even in the most challenging dataset ESPGame, the improvement of our method remains significant.
- Compared with the traditional non-DNN methods, DNN-based methods have greater advantages in this area. Moreover, in models based on DNN, the latter four methods in Table 2 that can address the challenge of missing views demonstrate more significant advantages. This further corroborates the necessity of designing dedicated methods specifically for incomplete data.

To investigate the impact of missing data, we conducted performance comparisons on Corel5k dataset at different missing percentages for views or labels, shown in Figure 2.

Hyper-parameter Analysis

In our tow-stage model, there are three hyper-parameters, *i.e.*, α, β, and γ that need to be set before training. In order to study the sensitivity of our model to the three hyper-parameters, we experiment on the Corel5k dataset and Pascal07 dataset with 50% available instances for each view, 50% missing labels, and 70% training samples. Figures 3a and 3b depict the AP values in relation to the hyper-parameters α and γ, while Figures 3c and 3d illustrate the AP curves concerning the selection of β. To ensure exper-



(a) missing-view study (b) missing-label study

Figure 2: Experimental results on the Corel5k dataset: (a) different missing-view rates combined with a 50% missing-label rate; (b) different missing-label rates combined with a 50% missing-view rate.

iment validity, irrelevant hyper-parameters were fixed. Obviously, when both of α and γ are correspondingly selected from the rages of $[1, 100]$ for Corel5k dataset and $[5, 500]$ of Pascal07 dataset, our model demonstrated consistent and satisfactory performance. As for β, Figures 3c and 3d reveal that our model's sensitivity to this parameter is not significant. We opted to set β within the range of $[1e-3, 1e-1]$.

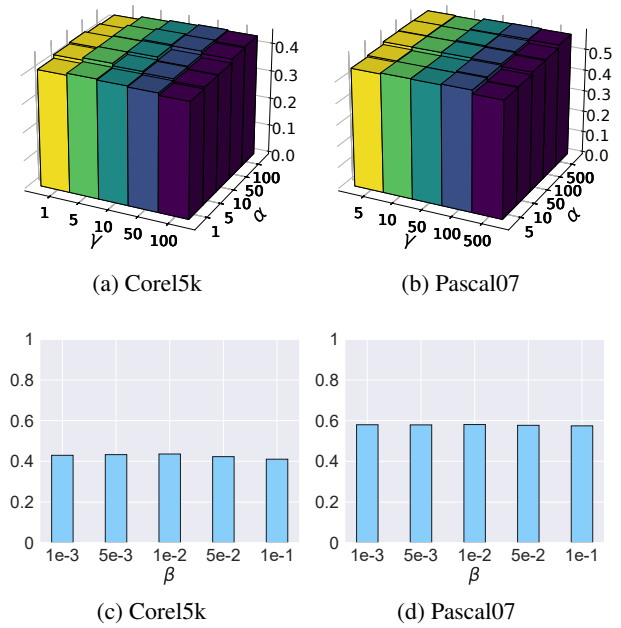
Ablation Study

To validate the efficacy of our approach, we conducted ablation experiments on the Corel5k and Pascal07 datasets, utilizing 50% of instances of each view, 50% with missing labels, and training with 70% of samples. We alter our loss function in several ways to test the necessity of each term of loss, namely $L_{IB}^{(k)}$, L_{CL1}, and L_{CL2}. The best results for each altered loss function, including the backbone L_{CE}, are presented in Table 3. Our observations are as follows: (i) The introduction of each loss component resulted in a enhancement in performance metrics. (ii) The most substantial improvement was associated with the inclusion of L_{CL1}.

Clearly, our two-stage model can be consolidated into a one-stage model by multiplying each $L_{IB}^{(k)}$ in the first stage by a common hyper-parameter and adding it to the second-stage loss L_{SCL}. However, we observed a performance degradation, shown in Tabel 3. We provide an intuitive explanation: When using the same hyper-parameter, the model fails to fully extract information from each view. For example, if $L_{IB}^{(1)}$ is larger, the model tends to prioritize optimizing $L_{IB}^{(1)}$, resulting in insufficient extraction of information from other views. One solution is to use m different hyper-parameters, but this also makes hyper-parameter tuning extremely challenging.

Backbone	$L_{IB}^{(k)}$	L_{CL1}	L_{CL2}	Corel5k AP	Corel5k AUC	Pascal07 AP	Pascal07 AUC
✓				0.375	0.882	0.547	0.847
✓	✓			0.383	0.882	0.547	0.848
✓		✓		0.389	0.906	0.559	0.855
✓			✓	0.376	0.885	0.551	0.847
✓	✓	✓		0.409	0.912	0.564	0.858
✓	✓		✓	0.395	0.907	0.549	0.848
✓		✓	✓	0.419	0.915	0.572	0.865
✓	✓	✓	✓	0.436	0.920	0.581	0.868
one stage				0.422	0.916	0.570	0.863

Table 3: The ablation experiment and one-stage experiment



(a) Corel5k (b) Pascal07

(c) Corel5k (d) Pascal07

Figure 3: AP values under different hyper-parameters: Different α and γ on the (a) Corel5k and (b) Pascal07 datasets, and different β on the (c) Corel5k and (d) Pascal07 datasets.

Conclusion

In this paper, we proposed a novel two-stage information extraction network to address the problem of iMvMLC. Unlike previous methods, the proposed method adopts an information-theoretic perspective to extract information from incomplete data. Extensive experimental evidence demonstrates the superiority of our method over existing state-of-the-art techniques, showcasing the promising prospects of utilizing information theory in addressing the iMvMLC problem.

Acknowledgments

This work is supported by Shenzhen Higher Education Stability Support Program Project under Grant No. GXWD20220811173317002 and National Natural Science Foundation of China under Grant No. 62372136.

References

Achille, A.; and Soatto, S. 2018. Information dropout: Learning optimal representations through noisy computa-

tion. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2897–2905.

Alemi, A.; Fischer, I.; Dillon, J.; and Murphy, K. 2017. Deep Variational Information Bottleneck. In *International Conference on Learning Representations*.

Amjad, R. A.; and Geiger, B. C. 2019. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE transactions on pattern analysis and machine intelligence*, 42(9): 2225–2239.

Belghazi, M. I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, D. 2018. Mutual information neural estimation. In *International Conference on Machine Learning*, 531–540. PMLR.

Duygulu, P.; Barnard, K.; de Freitas, J. F.; and Forsyth, D. A. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part IV 7*, 97–112. Springer.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Grubinger, M.; Clough, P.; Müller, H.; and Deselaers, T. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, volume 2.

Gutmann, M.; and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 297–304. JMLR Workshop and Conference Proceedings.

Huiskes, M. J.; and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 39–43.

Jiang, B.; Xiang, J.; Wu, X.; Wang, Y.; Chen, H.; Cao, W.; and Sheng, W. 2022. Robust multi-view learning via adaptive regression. *Information Sciences*, 610: 916–937.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 18661–18673.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

Lee, C.; and Van der Schaar, M. 2021. A variational information bottleneck approach to multi-omics data integration. In *International Conference on Artificial Intelligence and Statistics*, 1513–1521. PMLR.

Li, L.; Wan, Z.; and He, H. 2021. Incomplete multi-view clustering with joint partition and graph learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(1): 589–602.

Li, X.; and Chen, S. 2021. A concise yet effective model for non-aligned incomplete multi-view and missing multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 5918–5932.

Liu, C.; Wen, J.; Luo, X.; Huang, C.; Wu, Z.; and Xu, Y. 2023a. DICNet: Deep Instance-Level Contrastive Network for Double Incomplete Multi-View Multi-Label Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8807–8815.

Liu, C.; Wen, J.; Luo, X.; and Xu, Y. 2023b. Incomplete Multi-View Multi-Label Learning via Label-Guided Masked View- and Category-Aware Transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8816–8824.

Liu, C.; Wen, J.; Xu, Y.; Nie, L.; and Zhang, M. 2023c. Learning Reliable Representations for Incomplete Multi-View Partial Multi-Label Classification. *arXiv preprint arXiv:2303.17117*.

Liu, X.; Li, M.; Tang, C.; Xia, J.; Xiong, J.; Liu, L.; Kloft, M.; and Zhu, E. 2021. Efficient and Effective Regularized Incomplete Multi-View Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8): 2634–2646.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Shannon, C. E. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27: 379–423.

Soatto, S.; and Chiuso, A. 2014. Visual representations: Defining properties and deep approximations. *arXiv preprint arXiv:1411.7676*.

Sridharan, K.; and Kakade, S. M. 2008. An Information Theoretic Framework for Multi-view Learning. *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland*, 403–414.

Tan, Q.; Yu, G.; Domeniconi, C.; Wang, J.; and Zhang, Z. 2018. Incomplete multi-view weak-label learning. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 2703–2709.

Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; and Isola, P. 2020. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33: 6827–6839.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Trosten, D. J.; Lokse, S.; Jenssen, R.; and Kampffmeyer, M. 2021. Reconsidering representation alignment for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1255–1265.

Tsai, Y.-H. H.; Wu, Y.; Salakhutdinov, R.; and Morency, L.-P. 2020. Self-supervised Learning from a Multi-view Perspective. In *International Conference on Learning Representations*.

Von Ahn, L.; and Dabbish, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 319–326.

Wang, H.; Guo, X.; Deng, Z.-H.; and Lu, Y. 2022. Rethinking minimal sufficient representation in contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16041–16050.

Wang, Q.; Boudreau, C.; Luo, Q.; Tan, P.-N.; and Zhou, J. 2019. Deep multi-view information bottleneck. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, 37–45. SIAM.

Wang, S.; Liu, X.; Zhu, X.; Zhang, P.; Zhang, Y.; Gao, F.; and Zhu, E. 2021. Fast parameter-free multi-view subspace clustering with consensus anchor guidance. *IEEE Transactions on Image Processing*, 31: 556–568.

Wen, J.; Liu, C.; Deng, S.; Liu, Y.; Fei, L.; Yan, K.; and Xu, Y. 2023. Deep Double Incomplete Multi-View Multi-Label Learning With Incomplete Labels and Missing Views. *IEEE Transactions on Neural Networks and Learning Systems*.

Wen, J.; Zhang, Z.; Zhang, Z.; Wu, Z.; Fei, L.; Xu, Y.; and Zhang, B. 2020. DIMC-Net: Deep Incomplete Multi-View Clustering Network. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 3753–3761. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

Wu, X.; Chen, Q.-G.; Hu, Y.; Wang, D.; Chang, X.; Wang, X.; and Zhang, M.-L. 2019. Multi-view multi-label learning with view-specific information extraction. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 3884–3890.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3733–3742.

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

Xu, C.; Tao, D.; and Xu, C. 2015. Multi-View Learning With Incomplete Views. *IEEE Transactions on Image Processing*, 24(12): 5812–5825.

Xu, C.; Zhao, W.; Zhao, J.; Guan, Z.; Song, X.; and Li, J. 2022. Uncertainty-aware multiview deep learning for internet of things applications. *IEEE Transactions on Industrial Informatics*, 19(2): 1456–1466.

Zhang, C.; Yu, Z.; Hu, Q.; Zhu, P.; Liu, X.; and Wang, X. 2018. Latent semantic aware multi-view multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhang, M.-L.; and Zhou, Z.-H. 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8): 1819–1837.

Zhang, Q.; Yu, S.; Xin, J.; and Chen, B. 2022. Multi-view information bottleneck without variational approximation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4318–4322. IEEE.

Zhao, D.; Gao, Q.; Lu, Y.; Sun, D.; and Cheng, Y. 2021. Consistency and diversity neural network multi-view multi-label learning. *Knowledge-Based Systems*, 218: 106841.

Zhu, P.; Hu, Q.; Hu, Q.; Zhang, C.; and Feng, Z. 2018. Multi-view label embedding. *Pattern recognition*, 84: 126–135.