View-Category Interactive Sharing Transformer for Incomplete Multi-View Multi-Label Learning

Shilong Ou, Zhe Xue*, Yawen Li, Meiyu Liang, Yuanqiang Cai, Junjiang Wu
Beijing University of Posts and Telecommunications, China
{osl, xuezhe, meiyu1210, caiyuanqiang, wujunjiang}@bupt.edu.cn, warmly0716@126.com

Abstract

As a problem often encountered in real-world scenarios, multi-view multi-label learning has attracted considerable research attention. However, due to oversights in data collection and uncertainties in manual annotation, real-world data often suffer from incompleteness. Regrettably, most existing multi-view multi-label learning methods sidestep missing views and labels. Furthermore, they often neglect the potential of harnessing complementary information between views and labels, thus constraining their classification capabilities. To address these challenges, we propose a view-category interactive sharing transformer tailored for incomplete multi-view multi-label learning. Within this network, we incorporate a two-layer transformer module to characterize the interplay between views and labels. Additionally, to address view incompleteness, a KNN-style missing view generation module is employed. Finally, we introduce a view-category consistency guided embedding enhancement module to align different views and improve the discriminating power of the embeddings. Collectively, these modules synergistically integrate to classify the incomplete multi-view multi-label data effectively. Extensive experiments substantiate that our approach outperforms the existing state-of-the-art methods.

1. Introduction

The proliferation of data sources, coupled with advancements in data collection techniques, enables the extraction of diverse types of features, extending beyond a single perspective. This leads to an intriguing research question: how to effectively integrate multi-view features and extract meaningful features from them. Simultaneously, it is observed that single-label data often falls short in encapsulating the complexities of real-world scenarios. For example, an image of a flower may also warrant labels like 'leaf' or 'garden', illustrating that multi-labels provide a richer category feature space and more accurately preserve inter-label relationships [8]. Given its practical relevance, the field of multi-view multi-label learning emerges as a domain of substantial research interest.

In the field of multi-view learning, multi-view multi-label classification (MVMLC) poses a significant challenge. Considering the consistency across views and the complexity of labels, it is essential for methods to simultaneously align views and extract diverse features from an integrated representation. In recent years, numerous researchers have made strides towards this objective. One such effort is the introduction of a multi-view label embedding model [30], which employs the Hilbert-Schmidt Independence Criterion [4] to establish a link between view feature and category feature spaces. Concurrently, deep neural networks have been employed in this realm. A noteworthy development is the SIMM neural network [3], which utilizes adversarial and label losses to identify shared features while implementing regularization to capture view-specific details.

While MVMLC has attracted considerable research attention, it often operates under the presumption that all views and labels are complete. This assumption does not consistently align with real-world scenarios. For instance, a video might lack audio or text content. Multimedia data shared on social media networks might lack comprehensive annotations from users. Such instances culminate in the pervasive challenge of missing multi-view features and incomplete category information within real-world datasets. Some methods aim to address this problem by either masking or restoring the missing views [9, 20, 22, 25]. Although these approaches for handling incomplete multi-view or multi-label learning have achieved notable results, they often fall short in simultaneously addressing both types of incompleteness. To handle the double incomplete multi-view multi-label learning problem (DIMVMLC), several deep contrastive networks are proposed [10, 16]. Furthermore, an incomplete multi-view multi-label learning method based on transformers have been proposed [11], showing adaptability to multi-view and multi-label datasets.

However, most existing methods for double incomplete

*Corresponding author

multi-view multi-label classification overlook two critical aspects. Firstly, while some methods recognize the importance of complementary information across multiple views, they often underestimate the complementarity between multi-view features and categories. Efficiently facilitating interaction between views and categories can dramatically enhance the ability to counteract the simultaneous absence of both multi-view feature and category information. Secondly, these methods frequently overlook the issue of missing multi-view features and do not consider completing missing multi-view features. This negligence makes them susceptible to vulnerabilities in insufficient view features, thereby limiting their classification performance.

To overcome the prevailing challenges in double incomplete multi-view multi-label learning, we propose the View-category Interactive Sharing Transformer (**VIST**), as shown in Fig. 1. VIST is an architecturally advanced framework specifically designed to exploit the synergy between multiple views and categories and adeptly generate missing views. The dual capability of this method not only amplifies the effectiveness of multi-view embeddings through high-level semantics of categories but also significantly enhances the precision of multi-label classification by utilizing these enriched embeddings. This advancement inherently fortifies the method's resilience, particularly in scenarios characterized by the lack of specific views and labels, thereby ensuring robust and reliable performance. VIST integrates three key components: view-category interactive sharing transformer, missing view generation, and view-category consistency guided embedding enhancement. The view-category interactive sharing transformer, a compact two-layer network with four transformer blocks, fosters an interaction between multi-view and category information, uncovering their complementary potential. The missing view generation, adopting a KNN-style approach, adeptly completes missing views in line with each sample's distribution. The view-category consistency guided embedding enhancement module utilizes contrastive learning strategy to sharpen the discriminative strength of embeddings and boost classification accuracy. Extensive experiments demonstrate the excellent performance of VIST. In summary, this paper makes the following contributions:

- We introduce a novel view-category interactive sharing transformer for incomplete multi-view multi-label learning. Our method can effectively exploit the complementary information between views and categories through the interaction between multi-view feature and category information, thereby effectively counteracting the challenges posed by incomplete views and labels.
- We propose a missing view generation method specifically designed for the incomplete multi-view multi-label learning task. This method combines the KNN strategy with a multivariate Gaussian distribution, ensuring a statistically sound imputation of missing views that are cohesively aligned with the corresponding samples.
- To bolster the discriminating power of the multi-view embeddings, we develop a view-category consistency guided embedding enhancement module. It adopts contrastive learning to align embeddings across different views and leverages category information to guide embedding learning process, significantly boosting the effectiveness of the resulting embeddings and the performance of multi-label classification.

2. Related Work

Many methods have been proposed for the MVMLC task. lrMMC [12] ensures the common subspace remains low rank, making it compatible with matrix completion, while also learning combination weights to tap into the distinct strengths of each view. Matrix factorization is utilized to draw out complementary data across different views [27]. Despite their strengths, these models struggles with handling datasets that have missing views or labels. To solve this, MVL-IV [23] delves into incomplete multi-view learning by leveraging the relationships between views. The iMSF model is another incomplete multi-view single-label learning approach. It ingeniously splits the incomplete multi-view classification responsibilities into several complete subtasks [26]. However, both MVL-IV and iMSF have a common shortcoming: they only address view incompleteness. Conversely, MvEL targets the extraction of context and neighborhood consistency but is tailored only for the incomplete multi-label scenario [29].

On the other hand, for double incomplete multi-view multi-label learning task, iMvWL [14] and NAIML [9], have ventured into addressing the incompleteness in both views and labels. iMvWL aligns multi-view features and multi-label data into a shared subspace, enriched by a correlation matrix that bolsters the projection from label space to the embedding subspace. NAIML addresses this dual incompleteness by leaning on the low-rank assumption of the sub-label matrix, subtly harnessing sub-class correlations. In recent years, deep learning models have been employed to address this issue. DDINet [21], DICNet [10] and LMVCAT [11] are deep learning models for DIMVMLC. DDINet is the first deep learning model which contains a view-specific decoder network, successfully preserving the key information of raw views. DICNet is the first deep learning model for DIMVMLC which employs contrastive learning at the instance-level, guiding model to extract deep cross-view features. LMVCAT is a transformer-based double incomplete multi-view multi-label learning framework, consisting of two transformers specifically tailored for views and categories, and constructs a graph on incomplete labels to guide the encoder in extracting view-specific features.

3. Methodology

3.1. Problem Definition

Given n multi-view data points from V views, they can be represented as a set of matrices $\{\mathbf{X}_v\}_{v=1}^{V}$, with $\mathbf{X}_v \in \mathbb{R}^{n\times d_v}$. Here, d_v denotes the dimension of the v-th view. The label matrix for all samples is denoted by $\mathbf{Y} \in \{0,1\}^{n\times L}$, where L is the number of categories. If $\mathbf{Y}_{i,j} = 1$, it indicates that the i-th sample belongs to category j; if $\mathbf{Y}_{i,j} = 0$, it indicates non-membership in this category. To denote multi-view data with missing features, we introduce the missing-view indicator matrix $\mathbf{W} \in \{0,1\}^{n\times V}$. $\mathbf{W}_{i,j} = 1$ means that the j-th view of the i-th sample is available, whereas $\mathbf{W}_{i,j} = 0$ indicates unavailable. Similarly, we define the missing-label indicator matrix $\mathbf{U} \in \{0,1\}^{n\times L}$, where $\mathbf{U}_{i,j} = 1$ indicates that the j-th category of the i-th sample is available, and $\mathbf{U}_{i,j} = 0$ signifies the opposite. During the data pre-processing phase, all missing information for view $\mathbf{X}_v$ and label $\mathbf{Y}$ is assigned the value 0. The objective of our double incomplete multi-view multi-label learning task is to train a model capable of accurately predicting multiple categories for each sample. In subsequent sections, the detailed mechanisms of the proposed VIST are presented.

3.2. View-Category Interactive Sharing Transformer

Different views harbor complementary information, necessitating their interaction for effective utilization. Furthermore, category information embodies the data's high-level semantics, serving to augment the discriminative capability of multi-view representations. Inversely, multi-view features can remedy inaccuracies in label information comprehension due to incomplete labels. To this end, we introduce a view-category interactive sharing transformer. This novel transformer is designed to enhance information exchange between views and categories, consequently boosting the discriminative strength of multi-view representations and refining the precision of multi-label classification.

The view-category interactive sharing transformer comprises two key layers. The first layer focuses on extracting *shared features* that captures correlations among multiple views and categories. An essential preliminary step involves embedding the multi-view data $\mathbf{X}_v \in \mathbb{R}^{n\times d_v}$. This embedding process ensures the homogenization of feature dimensions across views. For a given sample $\mathbf{x}_v \in \mathbb{R}^{d_v}$ from $\mathbf{X}_v$, the embedding vector $\mathbf{e}_v \in \mathbb{R}^{d_e}$ can be denoted as $\mathbf{e}_v = Embedding(\mathbf{x}_v)$, where d_e denotes the dimension of embedding space, and $Embedding(\cdot)$ is a fully connected neural network. Then we stack the embedding vectors to obtain an original multi-view embedding sequence $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_V] \in \mathbb{R}^{V\times d_e}$, which is further used as input vectors for the transformer. Noted that for the incomplete multi-view data, we adopt missing view generation method (as detailed in Sec. 3.3) to produce embeddings of missing views, ensuring that $\mathbf{E}$ is complete across all views.

The structure of our transformer is the encoder module of the classical transformer [18]. A transformer block is denoted by $Transformer(\mathbf{\Phi})$, and the input of the transformer is $\mathbf{\Phi} \in \mathbb{R}^{t\times d_t}$, where t and d_t denote the number of tokens and the dimension of tokens. The first layer consists of two transformer blocks, denoted as $Transformer_{sv}$ and $Transformer_{sc}$. By discerning correlations among original views, $Transformer_{sv}$ serves to complement information across original views and get multi-view embedding $\tilde{\mathbf{E}}$, which is the enhanced representation of original views. It can be illustrated as follows:

$$\tilde{\mathbf{E}} = Transformer_{sv}(\mathbf{E}) \tag{1}$$

where $\tilde{\mathbf{E}} = [\tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2, \ldots, \tilde{\mathbf{e}}_V] \in \mathbb{R}^{V\times d_e}$. Here, an adaptive fusion layer is introduced to fuse the information from multiple views into a shared view feature s_v. The fusion process is formulated as follows:

$$\mathbf{s}_v = \sum_{v=1}^{V} \frac{e^{\theta_v^{\epsilon}} \tilde{\mathbf{e}}_v}{\sum_j e^{\theta_j^{\epsilon}}} \tag{2}$$

where $\tilde{\mathbf{e}}_v \in \mathbb{R}^{d_e}$ is the v-th row of $\tilde{\mathbf{E}}$, θ_v denotes the learnable weight and ϵ is a adjustment factor. By interacting with the shared view feature $\mathbf{s}_v$ and discerning correlations among original categories, $Transformer_{sc}$ serves to complement information across original categories and get category vectors $\tilde{\mathbf{C}}$, which can be illustrated as follows:

$$[\tilde{\mathbf{s}}_v, \tilde{\mathbf{C}}] = Transformer_{sc}([\mathbf{s}_v, \mathbf{C}]) \tag{3}$$

where $[\cdot,\cdot]$ denotes the concatenation operation and $\mathbf{C} \in \mathbb{R}^{L\times d_e}$ denotes original category vectors which are randomly initialized. Subsequently, the output feature of $Transformer_{sc}$, $\tilde{\mathbf{s}}_v$, is propagated into the second layer.

The second layer is designed to extract *advanced shared features*, which is achieved by promoting the interaction and fusion between view and category information based on the features extracted from the first layer, thereby obtaining a more discriminative multi-view data representation. This layer incorporates two transformer blocks, denoted as $Transformer_{av}$ and $Transformer_{ac}$. $Transformer_{av}$ is employed to complement information across views and extract advanced multi-view embedding $\bar{\mathbf{E}}$, which is the enhanced representation of views, by interacting with the shared category feature $\mathbf{s}_c$ and discerning view correlations, as represented below:

$$\mathbf{s}_c = Proj_{cv}(\tilde{\mathbf{s}}_v) \tag{4}$$

$$[\bar{\mathbf{s}}_c, \bar{\mathbf{E}}] = Transformer_{av}([\mathbf{s}_c, \tilde{\mathbf{E}}]) \tag{5}$$

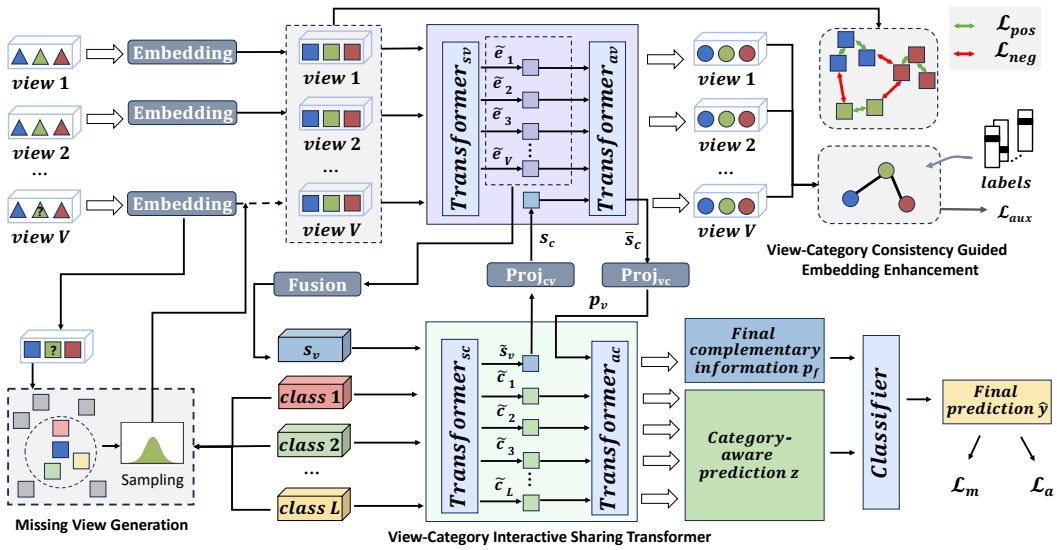


Figure 1. An overview of our method VIST. It comprises three modules, namely View-Category Interactive Sharing Transformer Module, Missing View Generation Module, and View-Category Consistency Guided Embedding Enhancement Module. Different colors represent different samples, while different shapes signify different features.

where $Proj_{cv}(\cdot)$ is a linear layer as a projection function designed to map vectors from the category feature space to the view feature space, $\bar{\mathbf{E}} = [\bar{\mathbf{e}}_1, \bar{\mathbf{e}}_2, \ldots, \bar{\mathbf{e}}_V] \in \mathbb{R}^{V \times d_e}$. Correspondingly, $Transformer_{ac}$ is employed to complement information across categories and extract advanced category vectors $\bar{\mathbf{C}}$ by leveraging the advanced shared feature $\mathbf{p}_v$ and discerning category correlations, as illustrated below:

$$\mathbf{p}_v = Proj_{vc}(\bar{\mathbf{s}}_c) \tag{6}$$

$$[\bar{\mathbf{p}}_v, \bar{\mathbf{C}}] = Transformer_{ac}([\mathbf{p}_v, \tilde{\mathbf{C}}]) \tag{7}$$

where $Proj_{vc}(\cdot)$ is a linear layer as a projection function to map vectors from the view feature space to the category feature space. Through the propagation of vectors $\mathbf{s}_v$, $\mathbf{s}_c$, and $\mathbf{p}_v$ among the transformer blocks, we facilitate the sharing of information between the view and category feature spaces, thereby extracting more refined and effective features of views and categories.

3.3. Missing View Generation

Considering the absence of views in the dataset, an intuitive consideration is that the proper completion of these missing views can enhance the model's performance. Through the contrastive learning process mentioned in Sec. 3.4, the distance between views of a sample tends to be relatively proximate. Thus, generating missing views utilizing the existing ones is a plausible approach. We adopt a method analogous to k-nearest neighbors, leveraging original category vectors $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_L]$ to assist in generating the missing views. Specifically, for a certain sample i, let $\mathcal{E} = \{v | \mathbf{W}_{i,v} = 1\}$ denote the index of existing views, and $\mathcal{M} = \{v | \mathbf{W}_{i,v} = 0\}$ denote the index of missing views. To complete the missing features of sample i with its original multi-view embedding $\mathbf{E}$, we first find the k nearest neighbors in the projected category feature space. The set of the neighbors $\mathcal{D}$ is constructed as follows:

$$\mathcal{D} = \{d | TopK(\sum_{j \in \mathcal{E}} \|\mathbf{e}_j - \mathbf{c}_d\|_2, d \in \{1, 2, 3, \ldots, L\})\} \tag{8}$$

where $TopK(\cdot)$ is a function designed to identify the indices of the top k categories, based on the smallest distance between embedding vectors and category vectors. Then, we employ a statistical method to describe the distribution of the missing views. We hypothesize that these missing views $\{\mathbf{e}_m\}_{m \in \mathcal{M}}$ adhere to a multivariate Gaussian distribution $\mathcal{N}(\mu, \mathbf{\Sigma})$, with a mean vector and covariance matrix denoted as:

$$\mu = \frac{\sum_{d \in \mathcal{D}} \mathbf{c}_d}{|\mathcal{D}|} \tag{9}$$

$$\mathbf{\Sigma} = \frac{1}{|\mathcal{D}| - 1} \sum_{d \in \mathcal{D}} (\mathbf{c}_d - \mu)(\mathbf{c}_d - \mu)^T \tag{10}$$

For the missing views, we sample from this distribution $|\mathcal{M}|$ times and substitute the missing views with the sampled results. Consequently, we can obtain the complete embeddings for the incomplete multi-view data. By reconstructing the absent multi-view data, our method further enhances its performance in category predictions.

3.4. View-Category Consistency Guided Embedding Enhancement

Through the aforementioned view-category interactive sharing transformer, we extract three multi-view embeddings $\mathbf{E}$, $\tilde{\mathbf{E}}$ and $\bar{\mathbf{E}}$ from different layers. In order to obtain

better classification performance, enhancing the discriminating power and effectiveness of these embeddings is of great importance. Specifically, predicated on the view consistency assumption [10, 24], the embedding $\mathbf{E}$ of a sample on different views should be aligned. Moreover, to improve the discriminating power of $\tilde{\mathbf{E}}$ and $\bar{\mathbf{E}}$, the consistency between category correlations and feature correlations of multi-view data can be leveraged. In light of these factors, we introduce view-category consistency guided embedding enhancement.

To learn more effective embedding $\mathbf{E}$, we use contrastive learning to align the embeddings of the same sample across different views. According to [13, 17], the loss function of contrastive learning can be formulated as follows:

$$\begin{aligned} &-\log \frac{e^{sim(x,x^+)}}{e^{sim(x,x^+)}+\sum_{x^-\in\mathcal{F}(x)} e^{sim(x,x^-)}} \\ &\approx -sim(x,x^+)+\log \sum_{x^-\in\mathcal{F}(x)} e^{sim(x,x^-)} \end{aligned} \tag{11}$$

where x, x^+, x^- denotes the anchor sample, its positive sample and negative sample. $\mathcal{F}(x)$ denotes the set of negative pairs. $sim(\cdot,\cdot)$ calculates the similarity between two samples.

Evidently, our objective is to maximize the similarity $sim(x,x^+)$ while simultaneously minimizing the similarity $\sum_{x^-\in\mathcal{F}(x)} e^{sim(x,x^-)}$. We denote $\mathbf{e}_{i,v}$ as the embedding of the v-th view for the i-th sample, then $\mathcal{F}(x)$ can be rewrited as $\mathcal{F}(i,v)=\{j,w|\sum^L \mathbf{y}_i\circ\mathbf{u}_i\circ\mathbf{y}_j\circ\mathbf{u}_j=0,\mathbf{U}_{j,w}=1\}$, where $\circ$ denotes Hadamard product, $\mathbf{u}_i$ denotes the missing vector of the i-th sample from U. As for maximizing $e^{sim(x,x^+)}$, we use the following loss function:

$$\mathcal{L}_{pos}=\sum_{v_1=1}^{V}\sum_{v_2=1,v_2\neq v_1}^{V}\frac{1}{d_e}\|n(\mathbf{e}_{i,v_1})-n(\mathbf{e}_{i,v_2})\|_2^2 \tag{12}$$

where $n(\cdot)$ denotes L_2 normalization function. To minimize $\sum_{x^-\in\mathcal{F}(x)} e^{sim(x,x^-)}$, the following loss function can be adopted:

$$\mathcal{L}_{neg}=\sum_{v_1=1}^{V}\sum_{j,v_2\in\mathcal{F}(i,v_1)}\frac{-1}{d_e}\|n(\mathbf{e}_{i,v_1})-n(\mathbf{e}_{j,v_2})\|_2^2 \tag{13}$$

To further improve the effectiveness of $\tilde{\mathbf{E}}$ and $\bar{\mathbf{E}}$, we utilize the category information to guide the learning process of these embeddings. We enhance the consistency of data correlations in the view feature space and the category feature space as detailed below:

$$\mathbf{G}_v^1(i,j)=(n(\tilde{\mathbf{e}}_{i,v})(n(\tilde{\mathbf{e}}_{j,v}))^T+1)/2 \tag{14}$$

$$\mathbf{G}_v^2(i,j)=(n(\bar{\mathbf{e}}_{i,v})(n(\bar{\mathbf{e}}_{j,v}))^T+1)/2 \tag{15}$$

$$\mathbf{H}=((\mathbf{Y}\circ\mathbf{U})(\mathbf{Y}\circ\mathbf{U})^T)./(\mathbf{U}\mathbf{U}^T) \tag{16}$$

where $\mathbf{G}_v^1$, $\mathbf{G}_v^2$ and $\mathbf{H}$ are the correlation matrices of $\tilde{\mathbf{E}}$, $\bar{\mathbf{E}}$ and labels respectively. $\mathbf{e}_{i,v}$ and $\bar{\mathbf{e}}_{i,v}$ represent the feature of the v-th view of the i-th sample from $\tilde{\mathbf{E}}$ and $\bar{\mathbf{E}}$ respectively. Through the cross-entropy loss function, we can align the data correlations across view feature space and category feature space so that more discriminative embedding can be obtained:

$$\begin{aligned} \mathcal{L}_{aux}=\frac{-1}{2nV}\sum_{v=1}^{V}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(&\mathbf{H}_{i,j}(\log\mathbf{G}_v^1(i,j)+\log\mathbf{G}_v^2(i,j)) \\ &+(\mathbf{U}_{i,j}-\mathbf{H}_{i,j})(\log(1-\mathbf{G}_v^1(i,j))+\log(1-\mathbf{G}_v^2(i,j)))) \end{aligned} \tag{17}$$

Considering the above factors, the loss function for view-category consistency guided embedding enhancement can be summarized as:

$$\mathcal{L}_c=\mathcal{L}_{pos}+\mathcal{L}_{neg}+\alpha\mathcal{L}_{aux} \tag{18}$$

where α is a weight parameter. Through optimizing the above loss function, the learned data embeddings become more discriminative, laying a solid foundation for subsequent multi-label classification.

3.5. Overall Loss Function

To achieve multi-label classification results, we partition the output $[\bar{\mathbf{p}}_v,\bar{\mathbf{C}}]$ of $Transformer_{ac}$ into $\bar{\mathbf{p}}_v$ and $\bar{\mathbf{C}}$. Specifically, for sample i, $\bar{\mathbf{p}}_v^i$ passes through a linear layer to yield $\mathbf{p}_f^i\in[0,1]^L$, representing the final complementary information between views and categories. $\bar{\mathbf{C}}^i$ is processed through a set of category-aware linear layers, resulting in $\mathbf{z}^i\in[0,1]^L$, which signifies category-aware predictions. With the assistance of $\mathbf{p}_f^i$, we get the final prediction $\hat{\mathbf{y}}^i$:

$$\hat{\mathbf{y}}^i=classifier(\mathbf{p}_f^i+\mathbf{z}^i) \tag{19}$$

where $classifier(\cdot)$ is a linear classifier. We utilize the masked cross-entropy function to compute the multi-label classification loss:

$$\mathcal{L}_m=\frac{\sum_{i=1}^{n}\sum_{l=1}^{L}(\mathbf{Y}_{i,l}\log\hat{\mathbf{y}}_l^i+(1-\mathbf{Y}_{i,l})\log(1-\hat{\mathbf{y}}_l^i))\mathbf{U}_{i,l}}{-\sum_{i,l}\mathbf{U}_{i,l}} \tag{20}$$

It is important to note that class imbalance is an inherent characteristic of multi-label data, impacting the generalization performance of multi-label prediction [15, 28]. Typically, for certain class labels, the instances of positive label assignments in training data are significantly fewer than those of negative assignments. To address this class-imbalance issue, we introduce a masked asymmetric loss function:

$$\mathbf{L}_l^i=(\max(\hat{\mathbf{y}}_l^i-0.5,0))^2\log(1-\max(\hat{\mathbf{y}}_l^i-0.5,0)) \tag{21}$$

$$\mathcal{L}_a = \frac{\sum_{i=1}^{n}\sum_{l=1}^{L}(\mathbf{Y}_{i,l}\log\hat{\mathbf{y}}_l^i + (1-\mathbf{Y}_{i,l})\mathbf{L}_l^i)\mathbf{U}_{i,l}}{-\sum_{i,l}\mathbf{U}_{i,l}} \quad (22)$$

Ultimately, we derive the overall loss function of the model as follows:

$$\mathcal{L}_o = \mathcal{L}_m + \beta\mathcal{L}_a + \gamma\mathcal{L}_c \quad (23)$$

where β, γ denotes different weight parameters. By optimizing $\mathcal{L}_o$, our method proficiently learns the embeddings for incomplete multi-view data and consequently produces accurate multi-label classification results.

4. Experiments

4.1. Experimental Setting

4.1.1 Datasets

We follow [9, 11, 14] and evaluate the performance of the proposed VIST on five datasets. These datasets encompass:

- **Corel 5k** [1]: The Corel5k dataset is a standard dataset frequently utilized in image annotation and retrieval research, which has 5000 images with 260 types of tags.
- **Pascal07** [2]: Pascal07, also known as PASCAL VOC 2007, is a renowned dataset for visual object classes, instrumental in object recognition and classification. It consists of 9963 images and 20 types of tags.
- **Espgame** [19]: The Espgame dataset is derived from an online game where participants generate labels for images, which contains 20770 samples and 268 types of tags.
- **IAPRTC12** [5]: IAPRTC12 is an international image annotation dataset, comprising approximately 20,000 images and 290 categories. These images cover a wide range of subjects and scenes, each accompanied by detailed descriptive text.
- **Mirflickr** [7]: The Mirflickr dataset consists of 25,000 images and 38 types of tags sourced from Flickr. Selected from the public domain, these images include a variety of themes and scenes.

We adopt the multi-view feature extraction method outlined in [6]. The images in these datasets undergo a comprehensive preprocessing procedure that transforms them into six distinct views. These views are GIST, HSV, HUE, LAB, RGB, and SIFT. Each of these views offers a unique perspective on the image data, capturing different aspects of visual information.

4.1.2 Implementation Details

For datasets, in line with [14], we process the five datasets to emulate an incomplete scenario as described below: For every view, we randomly exclude 50% of the samples, ensuring that each sample retains at least one view. Within each category, we arbitrarily designate 50% of both positive and negative tags as missing labels. For parameter settings, we choose $d_e = 512$, $\alpha = 0.1$, $\beta = 0.1$, $\gamma = 0.1$ and $k = 20$ for all the experiments. The model are trained for 500 epochs with a batch-size of 128 via SGD optimizer on an NVIDIA RTX A6000 GPU. To ensure the credibility of the results, we repeat the experiments 10 times and record the mean and variance of the outcomes.

4.1.3 Compared Methods

We choose ten representative methods for comparison. Out of these, eight methods - lrMMC, MVL-IV, iMSF, iMvWL, NAIML, DDINet, DICNet and LMVCAT - are detailed in Sec. 2. Additionally, we incorporate GLOCAL [31], to broaden the assessment spectrum. Because iMvWL and NAIML cater to datasets with missing views and labels, we have to modify other methods for consistency. Drawing inspiration from [14] and [9], we input average values from accessible views into lrMMC for the non-operational ones. For MVL-IV and iMSF, absent tags are treated as negative. GLOCAL are executed individually for each view, and we highlight their optimal outcomes. We adhere to the recommended settings for these comparative techniques, as mentioned in their respective publications or code repositories, to maintain an impartial evaluation.

4.1.4 Evaluation Metrics

Similar to [11], we employ Average Precision (AP), Ranking Loss (RL) and Area Under Curve (AUC). For a convenient comparison, we adopt 1-RL for evaluation. A larger value indicates superior model performance.

4.2. Experimental Results

In Tab. 1, we show the experimental results of all the methods on five datasets with 70% training samples, 50% missing views and missing labels. Based on the results from Tab. 1, our method exhibits a distinct advantage across most metrics. For instance, on the AP metric for dataset Corel5k, our method surpasses the second-best method, LMVCAT, by 3%. Besides, the performance of DDINet, DICNet and LMVCAT surpasses that of the first six methods. This can be attributed to their applicability specifically to the DIMVMLC issue. Additionally, they are both deep learning models, their superior outcomes underscore the vast potential deep learning holds for addressing this particular challenge.

While our method primarily addresses the DIMVMLC issue, we also evaluate the model's performance in the full-view and full-label context. The results are presented in Tab. 2. It can be observed that the model still exhibits superior performance in this scenario, suggesting that our method can be effectively applied in MVMLC.

Table 1. Experimental results of different methods on the five datasets with 70% training samples, 50% missing instances and missing labels. The best results are marked in bold. For simplicity, percent sign is omitted in the following section.

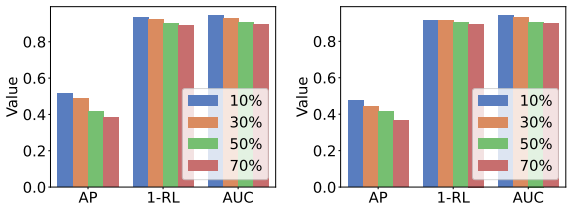
Dataset	Metric	lrMMC	MVL-IV	iMSF	GLOCAL	iMvWL	NAIML	DDINet	DICNet	LMVCAT	VIST
Corel5k	AP	24.0±0.2	24.0±0.1	18.9±0.2	28.5±0.4	28.3±0.7	30.9±0.4	36.4±0.1	38.1±0.4	38.4±0.4	**41.5±0.2**
	1-RL	76.2±0.2	75.6±0.1	70.9±0.5	80.4±0.3	86.5±0.3	87.8±0.2	87.1±0.0	88.2±0.4	88.0±0.2	**90.2±0.2**
	AUC	76.3±0.2	76.2±0.1	66.3±0.5	84.3±0.3	86.8±0.3	88.1±0.2	87.5±0.1	88.4±0.4	88.3±0.2	**90.6±0.1**
Pascal07	AP	42.5±0.3	43.3±0.2	32.5±0.0	49.6±0.4	44.1±1.7	48.8±0.3	53.6±0.2	50.5±1.2	51.9±0.5	**53.9±0.3**
	1-RL	69.8±0.3	70.2±0.1	56.8±0.0	76.7±0.4	73.7±0.9	78.3±0.1	80.7±0.1	78.3±0.8	81.1±0.4	**85.6±0.1**
	AUC	72.8±0.2	73.0±0.1	68.6±0.5	62.0±0.1	78.6±0.3	76.7±1.2	82.7±0.0	87.6±0.2	83.4±0.4	**88.3±0.2**
Espgame	AP	18.8±0.0	18.9±0.0	10.8±0.0	22.1±0.2	24.2±0.3	24.6±0.2	28.3±0.1	29.7±0.2	29.4±0.4	**30.7±0.3**
	1-RL	77.7±0.1	77.8±0.0	72.2±0.2	78.0±0.4	80.7±0.1	81.8±0.2	81.5±0.0	83.2±0.1	82.8±0.2	**84.4±0.0**
	AUC	78.3±0.1	78.4±0.1	67.4±0.3	78.4±0.4	81.3±0.2	82.4±0.2	82.0±0.1	83.2±0.1	82.8±0.2	**85.0±0.1**
IAPRTC12	AP	19.7±0.0	19.8±0.0	10.1±0.0	25.6±0.2	23.5±0.4	26.1±0.1	30.3±0.2	32.3±0.1	31.7±0.3	**33.9±0.2**
	1-RL	80.1±0.0	79.9±0.1	63.1±0.0	82.5±0.2	83.3±0.3	84.8±0.1	85.3±0.1	87.3±0.1	87.0±0.1	**88.4±0.2**
	AUC	80.5±0.0	80.4±0.1	66.5±0.1	83.0±0.1	83.6±0.2	85.0±0.1	85.4±0.0	87.4±0.1	87.2±0.1	**88.6±0.1**
Mirflickr	AP	44.1±0.1	44.9±0.1	32.3±0.0	53.7±0.2	49.5±1.2	55.1±0.2	59.8±0.2	58.9±0.5	59.4±0.5	**60.4±0.3**
	1-RL	80.5±0.0	80.4±0.1	66.5±0.1	83.2±0.1	83.6±0.2	85.0±0.1	86.3±0.0	86.3±0.4	86.5±0.3	**87.9±0.1**
	AUC	80.6±0.1	80.7±0.0	76.1±0.1	82.8±0.1	79.4±1.5	83.7±0.1	85.2±0.1	84.9±0.4	85.3±0.3	**86.9±0.1**

Table 2. Experimental results of different methods on the five datasets with 70% training samples, full views and labels. The best results are marked in bold.

Dataset	Metric	NAIML	DICNet	LMVCAT	VIST
Corel5k	AP	32.7	50.9	52.1	**55.4**
	1-RL	89.0	92.9	92.8	**93.5**
	AUC	89.3	93.1	93.0	**94.2**
Pascal07	AP	49.6	60.8	62.9	**65.5**
	1-RL	79.5	85.9	87.8	**89.4**
	AUC	82.2	87.6	89.2	**91.2**
Espgame	AP	25.1	39.1	38.5	**40.1**
	1-RL	82.5	87.1	87.6	**87.7**
	AUC	83.0	87.4	88.0	**89.6**
IAPRTC12	AP	26.7	41.8	43.6	**45.7**
	1-RL	82.5	91.2	91.8	**92.7**
	AUC	83.0	91.1	91.8	**92.0**
Mirflickr	AP	55.5	65.9	68.4	**70.4**
	1-RL	84.7	89.5	90.5	**91.3**
	AUC	83.9	87.6	88.9	**89.3**

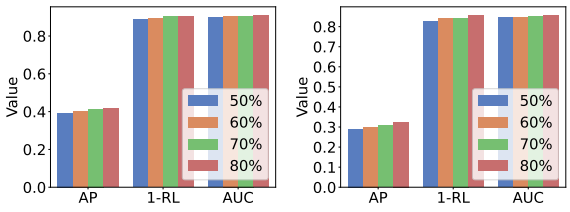
Fig. 2 illustrates the performance of the model under various missing ratios. It is interesting that while both incomplete views and incomplete labels detrimentally affect the model's performance, the latter has a more pronounced impact. Moreover, despite our efforts in view imputation, the model's accuracy noticeably diminishes as the view missing ratio escalates. This is primarily due to our reliance on existing views for imputing the missing views. As the number of missing views increases, the imputation algorithm tends to degrade towards random generation.

Fig. 3 depicts the performance of the model under various training sample ratios. While the ratio's impact on the model is not markedly significant, a general trend is observed: as the ratio increases, there is a corresponding improvement in the model's performance. Thus, selecting a 70% training sample ratio for our experiments is a proper choice.



(a) Missing-View Ratios (b) Missing-Label Ratios

Figure 2. The results on the Corel5k dataset are presented with (a) different missing-view ratios accompanied by a 50% missing-label ratio and (b) a consistent 50% missing-view ratio paired with different missing-label ratios.

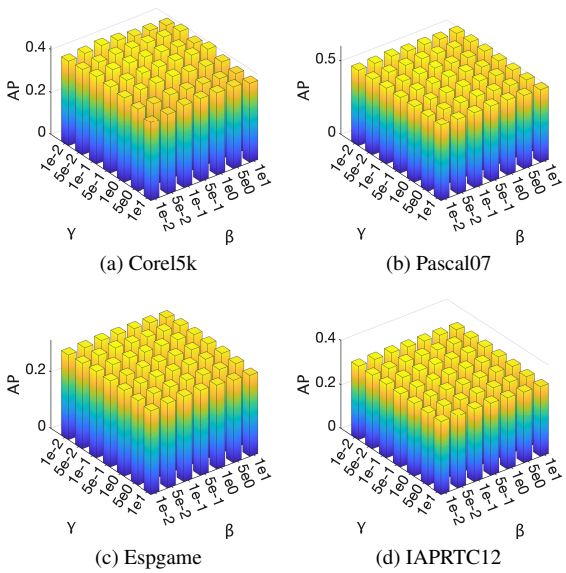


(a) Corel5k (b) Espgame

Figure 3. The results of different training sample ratios on (a) Corel5k dataset and (b) Espgame dataset with 50% missing-view ratio and 50% missing-label ratio.

4.3. Parameter Analysis

We evaluate four hyperparameters in our model, i.e, α, β, γ and k. To assess the influence of these hyperparameters on model performance, we adjust their values within their respective ranges and test their combinations based on the AP metric on different dataset with 50% missing views, 50% missing labels and 70% training sample ratio. Note that k is inherently related to the number of labels, thus, the actual value of $\hat{k}$ utilized in computation should be determined by $\hat{k}_j = \frac{k_i}{L_i} L_j$, where i and j represents the i-th



(a) Corel5k (b) Pascal07

(c) Espgame (d) IAPRTC12

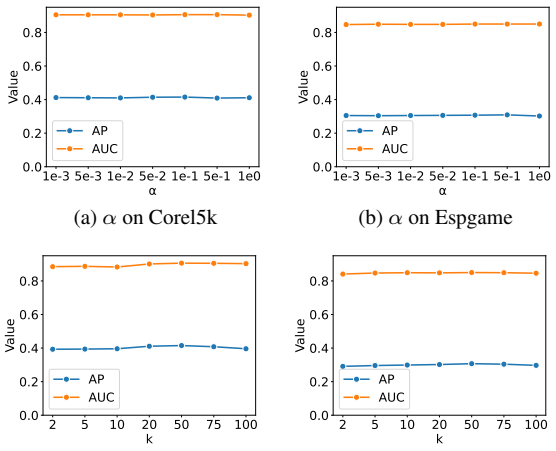
Figure 4. The results of different value combinations of hyperparameter β and γ on different datasets. α and k are respectively set to default values of 0.1 and 20.

and j-th dataset. Based on Fig. 4, β and γ exhibit comparatively favorable performances within the ranges of [5e-2, 5e-1] and [5e-2, 1e0] respectively on dataset Corel5k. This indicates that the values of β and γ should neither be excessively large nor too small, and the optimal value of β approximated around 0.1, the optimal value of γ should also be selected in the vicinity of 0.1.

As for α and k, we assess the impact of varying α and k on the model under fixed conditions of $\beta = 0.1$ and $\gamma = 0.1$ by measuring metrics AP and AUC. As illustrated in Fig. 5, on datasets Corel5k and Espgame, both the AP and AUC metrics remain within a narrow range, indicating that the model is not highly sensitive to α and k. This might be because they operate on the deep-layer mechanism, variations within certain ranges do not significantly impact the model's performance.

4.4. Ablation Study

In the ablation study, we evaluate the performance of each module of VIST on three representative datasets, the results are shown in Tab. 3. We can observe that view-category interactive sharing transformer module plays a pivotal role in enhancing the model's performance. Furthermore, the view-category consistency guided embedding enhancement module is able to improve the discriminating power of embeddings and boost the classification performance. The missing view generation module also contributes to the model's overall efficacy. Additionally, upon the removal of $\mathcal{L}_a$, there is a noticeable decline in the model's metrics, particularly in the AUC. This suggests that the masked asymmetric loss indeed addresses issues associated with imbalanced data. The ablation study demonstrates the effectiveness and reasonableness of each proposed module.



(a) α on Corel5k (b) α on Espgame

(c) k on Corel5k (d) k on Espgame

Figure 5. The results of different value of hyperparameter α and k on different datasets. β and γ are respectively set to default values of 0.1 and 0.1.

Table 3. Results of ablation experiments on the three representative datasets. Vanilla Transformer denotes the encoders of four vanilla transformers, $\mathbf{M}_1$ represents base view-category interactive sharing transformer module, $\mathbf{M}_2$ signifies the view-category consistency guided embedding enhancement module and $\mathbf{M}_3$ denotes missing view generation module.

Method	Corel5k		Espgame		IAPRTC12	
	AP	AUC	AP	AUC	AP	AUC
Vanilla Transformer	34.9	87.8	27.6	83.4	28.1	87.2
$\mathbf{M}_1$	39.5	89.6	29.4	84.1	31.6	87.8
$\mathbf{M}_1 + \mathbf{M}_2$	40.4	90.1	29.6	84.7	32.7	88.3
$\mathbf{M}_1 + \mathbf{M}_2 + \mathbf{M}_3$	**41.5**	**90.2**	**30.7**	**85.0**	**33.9**	**88.6**
w/o $\mathcal{L}_a$	40.6	89.8	29.6	84.5	32.4	88.2

5. Conclusion

In this paper, we propose a novel view-category interactive sharing transformer VIST for addressing the DIMVMLC challenge. Our method leverages a two-layer view-category transformer architecture to extract deep representations of both views and labels. Additionally, we employ a KNN-style approach with the multivariate Gaussian distribution to complete the missing views. To further enhance the efficacy of our model, a view-category consistency guided embedding enhancement module is integrated, significantly boosting the embedding quality and classification accuracy. Our comprehensive experimental analysis convincingly demonstrates that our method outperforms existing state-of-the-art approaches.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.62272058, No.62192784, No.62172056), and Beijing Natural Science Foundation (No.4242027, No.4232025).

References

[1] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Computer Vision — ECCV 2002*, pages 97–112, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. 6

[2] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 6

[3] Zheng Fang and Zhongfei Zhang. Simultaneously combining multi-view multi-label learning with maximum margin classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 864–869. IEEE, 2012. 1

[4] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005. 1

[5] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The iapr tc12 benchmark: A new evaluation resource for visual information systems. *Workshop Ontoimage*, 2006. 6

[6] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification. In *2010 IEEE Computer society conference on computer vision and pattern recognition*, pages 902–909. IEEE, 2010. 6

[7] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43, 2008. 6

[8] Junlong Li, Peipei Li, Xuegang Hu, and Kui Yu. Learning common and label-specific features for multi-label classification with correlation information. *Pattern Recognition*, 121:108259, 2022. 1

[9] Xiang Li and Songcan Chen. 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, 2022. 1, 2, 6

[10] Chengliang Liu, Jie Wen, Xiaoling Luo, Chao Huang, Zhihao Wu, and Yong Xu. Dicnet: Deep instance-level contrastive network for double incomplete multi-view multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8807–8815, 2023. 1, 2, 5

[11] Chengliang Liu, Jie Wen, Xiaoling Luo, and Yong Xu. Incomplete multi-view multi-label learning via label-guided masked view- and category-aware transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8816–8824, 2023. 1, 2, 6

[12] Meng Liu, Yong Luo, Dacheng Tao, Chao Xu, and Yonggang Wen. Low-rank multi-view learning in matrix completion for multi-label image classification. In *Proceedings of the AAAI conference on artificial intelligence*, 2015. 2

[13] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*, 2018. 5

[14] Qiaoyu Tan, Guoxian Yu, Carlotta Domeniconi, Jun Wang, and Zili Zhang. Incomplete multi-view weak-label learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 2703–2709. International Joint Conferences on Artificial Intelligence Organization, 2018. 2, 6

[15] Adane Nega Tarekegn, Mario Giacobini, and Krzysztof Michalak. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965, 2021. 5

[16] Daniel J Trosten, Sigurd Lokse, Robert Jenssen, and Michael Kampffmeyer. Reconsidering representation alignment for multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1255–1265, 2021. 1

[17] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 5

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 3

[19] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, page 319–326, New York, NY, USA, 2004. Association for Computing Machinery. 6

[20] Jie Wen, Zheng Zhang, Lunke Fei, Bob Zhang, Yong Xu, Zhao Zhang, and Jinxing Li. A survey on incomplete multiview clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(2):1136–1149, 2022. 1

[21] Jie Wen, Chengliang Liu, Shijie Deng, Yicheng Liu, Lunke Fei, Ke Yan, and Yong Xu. Deep double incomplete multi-view multi-label learning with incomplete labels and missing views. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2

[22] Mengyao Xie, Zongbo Han, Changqing Zhang, Yichen Bai, and Qinghua Hu. Exploring and exploiting uncertainty for incomplete multi-view classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19873–19882, 2023. 1

[23] Chang Xu, Dacheng Tao, and Chao Xu. Multi-view learning with incomplete views. *IEEE Transactions on Image Processing*, 24(12):5812–5825, 2015. 2

[24] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16051–16060, 2022. 5

[25] Nan Xu, Yanqing Guo, Xin Zheng, Qianyu Wang, and Xiangyang Luo. Partial multi-view subspace clustering. In *Proceedings of the 26th ACM International conference on multimedia*, pages 1794–1801, 2018. 1

[26] Lei Yuan, Yalin Wang, Paul M Thompson, Vaibhav A Narayan, Jieping Ye, Alzheimer's Disease Neuroimaging Initiative, et al. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *NeuroImage*, 61(3):622–632, 2012. 2

[27] Changqing Zhang, Ziwei Yu, Qinghua Hu, Pengfei Zhu, Xinwang Liu, and Xiaobo Wang. Latent semantic aware multi-view multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 2

[28] Min-Ling Zhang, Yu-Kun Li, Hao Yang, and Xu-Ying Liu. Towards class-imbalance aware multi-label learning. *IEEE Transactions on Cybernetics*, 52(6):4459–4471, 2022. 5

[29] Wei Zhang, Ke Zhang, Pan Gu, and Xiangyang Xue. Multi-view embedding learning for incompletely labeled data. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013. 2

[30] Pengfei Zhu, Qi Hu, Qinghua Hu, Changqing Zhang, and Zhizhao Feng. Multi-view label embedding. *Pattern recognition*, 84:126–135, 2018. 1

[31] Yue Zhu, James T Kwok, and Zhi-Hua Zhou. Multi-label learning with global and local label correlation. *IEEE Transactions on Knowledge and Data Engineering*, 30(6):1081–1094, 2017. 6