# Incomplete Multi-View Multi-Label Learning via Label-Guided Masked View- and Category-Aware Transformers

**Chengliang Liu**[1], **Jie Wen**[1*], **Xiaoling Luo**[1], **Yong Xu**[1,2*]

[1] Shenzhen Key Laboratory of Visual Object Detection and Recognition, Harbin Institute of Technology, Shenzhen, China
[2] Pengcheng Laboratory, Shenzhen, China.
liucl1996@163.com, jiewen_pr@126.com, xiaolingluoo@outlook.com, yongxu@ymail.com

## Abstract

As we all know, multi-view data is more expressive than single-view data and multi-label annotation enjoys richer supervision information than single-label, which makes multi-view multi-label learning widely applicable for various pattern recognition tasks. In this complex representation learning problem, three main challenges can be characterized as follows: i) How to learn consistent representations of samples across all views? ii) How to exploit and utilize category correlations of multi-label to guide inference? iii) How to avoid the negative impact resulting from the incompleteness of views or labels? To cope with these problems, we propose a general multi-view multi-label learning framework named label-guided masked view- and category-aware transformers in this paper. First, we design two transformer-style based modules for cross-view features aggregation and multi-label classification, respectively. The former aggregates information from different views in the process of extracting view-specific features, and the latter learns subcategory embedding to improve classification performance. Second, considering the imbalance of expressive power among views, an adaptively weighted view fusion module is proposed to obtain view-consistent embedding features. Third, we impose a label manifold constraint in sample-level representation learning to maximize the utilization of supervised information. Last but not least, all the modules are designed under the premise of incomplete views and labels, which makes our method adaptable to arbitrary multi-view and multi-label data. Extensive experiments on five datasets confirm that our method has clear advantages over other state-of-the-art methods.

## Introduction

Since the development of representation learning methods, data analysis technology based on simple single-view data has become more and more difficult to meet diverse application requirements. In the past few years, data acquisition technology has flourished, and multi-view data from various media or with different styles are ubiquitous, providing more possibilities for comprehensively and accurately describing observation targets. Simply put, multiple observations, obtained from different observation angles for the same thing, could be regarded as multi-view data (Li and He 2020; Li, Wan, and He 2021; Wang et al. 2022). For example, retinal images captured at four different positions constitute a four-view retinal dataset (Luo et al. 2021; Wen et al. 2022; Luo et al. 2023); the features extracted from the target image with different feature extraction operators can also form a multi-view dataset (Liu et al. 2022; Hu, Lou, and Ye 2021). More typically, multi-view or multi-modal datasets composed of multimedia data such as text, pictures, and videos have been widely used in many applications such as web page classification (Hu, Shi, and Ye 2020; Lu et al. 2019; Liu et al. 2023). As a result, multi-view learning emerges as the times require, and a large number of methods based on subspace learning, matrix factorization, and graph learning have been proposed (Zhan et al. 2017; Liu et al. 2017; Wen et al. 2019). Most of these methods seek to obtain a consistent representation of multiple views to characterize the essential attributes of objects. On the other hand, since the core of multi-view learning is representation learning, researchers usually combine it with downstream tasks to improve the application value and evaluate the learning effect (Chen, Wang, and Lai 2022). That is, multi-view learning can be divided into clustering and classification, according to whether supervised information is available. Further, in the single-label classification task, a sample is only labeled with one category, which is obviously against the distribution of information in nature (Zhao et al. 2022). For example, a bird picture is likely to contain the category of 'sky' or 'tree'. Therefore, although multi-label classification still faces more challenges than single-label classification, its broad application prospects are attracting a lot of research enthusiasm. Based on this, a complex fusion task, *i.e.*, multi-view multi-label classification (*MvMlC*) is put on the agenda. Zhang et al. proposed a matrix factorization-based *MvMlC* model, which attempts to enforce alignment of different views in the kernel space to exploit multi-view complementarity and explore more latent information (Zhang et al. 2018). A Bernoulli mixture conditional model is proposed to model label dependencies and employ a variational inference framework for approximate posterior estimation (Sun and Zong 2020).

Researchers have conducted extensive researches in the field of *MvMlC*, however these works assume that all views and labels are complete, which is often violated in practice. For example, if a web page contains only text and images, then the video view of the page is not available. To avoid

*Corresponding authors.

the negative effects of missing views as much as possible, some multi-view learning works try to mask unavailable views or restore missing views (Wen et al. 2022; Xu et al. 2018). Similarly, manual annotation is likely to miss some tags due to mistakes or cost, which inevitably weakens the supervision information of multi-label. To solve the issue, some works focusing on incomplete multi-label classification have been developed in recent years (Zhu, Kwok, and Zhou 2018; Huang et al. 2019). Although these methods designed for incomplete multi-view or incomplete multi-label learning have achieved surprising fruits, most of them are not able to cope with the both incomplete cases simultaneously.

To this end, we propose a general *MvMlC* framework, termed **L**abel-guided **M**asked **V**iew- and **C**ategory-**A**ware **T**ransformers (**LMVCAT**), which can handle the multi-view data with arbitrary missing-view and missing-label. In addition, unlike traditional methods to learn low-level representations of samples, deep neural networks based LMVCAT is capable to extract high-level features of samples, which is of great benefit for learning complex multi-category semantic labels. In recent years, transformer, designed for natural language processing, has shown its dominance in the image and other fields, which clearly demonstrates the effectiveness of the self-attention mechanism (Vaswani et al. 2017; Huang et al. 2022b,a). Inspired by this, our LMVCAT is also designed based on the transformer with self-attention mechanism which can provide a global receptive field for each view (Lanchantin et al. 2021). Overall, our model is composed of four parts: a masked view-aware encoder, an adaptively weighted multi-view fusion module, a label-guided sample-level graph constraint, and a subcategory-aware multi-label classification module. Specifically, the four parts of our LMVCAT are motivated by the following points: 1) breaking the inter-view barriers to aggregate multi-view features can fully exploit the complementary information of multiple views, so we design a view-aware module to integrate all views while extracting sample-specific high-level representations; 2) different discrimination ability of views means the different contributions to the final classification, so an adaptively weighted strategy that automatically learns the weight factor of each view is needed (Chen et al. 2022); 3) compared with single-label data, multi-label data naturally enjoys richer label similarity information, so it is of great significance to utilize label smoothness (manifold) to guide sample encoding; 4) it is well known that multi-label data is with non-negligible inter-class correlations, thus we adopt a category-aware module that learns correlations in the subcategory embedding space to collaboratively predict labels (Lanchantin et al. 2021). Apart from those, it needs to be emphasized that we consider the possibility of missing labels and missing views in all components of our model. So it is no doubt that our LMVCAT is a general *MvMlC* framework that is comfortable with all kinds of multi-view multi-label data. Our contributions are summarized as follows:

- To the best of our knowledge, this is the first Transformer-based incomplete multi-view multi-label learning framework capable of handling both incomplete views and labels. The proposed masked view-aware self-attention module could avoid the missing views' negative effects to information interaction across views. Similarly, our category-aware module is designed to mine latent inter-class correlations in the subcategory embedding space to improve the expressiveness.
- Label manifold is fully exploited. Although the multi-label information is fragmentary, we still try our best to construct an approximate similarity graph based on incomplete labels to guide the encoding process of samples, which further strengthens the discrimination power of high-level semantic features.
- Different views contribute different importance to the prediction, so we introduce an adaptively weighted fusion method to balance this importance instead of simply adding multiple views. Sufficient experimental results confirm the effectiveness of our LMVCAT.

# Preliminaries

## Formulation

In this section, we define our main problem as follows: Given input multi-view dataset $\mathbf{D} = \{\mathbf{X}, \mathbf{Y}\}$, which contains $n$ samples. For $i$-th sample, $\mathbf{X}_i$ is composed of $m$ views with dimension $d_v$, *i.e.*, $\mathbf{X}_i = \left\{x_i^{(v)} \in \mathbb{R}^{d_v}\right\}_{v=1}^{m}$, or for $v$-th view, $\mathbf{X}^{(v)} = \left\{x_i^{(v)} \in \mathbb{R}^{d_v}\right\}_{i=1}^{n}$. $\mathbf{Y}_i \in \{0,1\}^c$ is a row vector that denotes the label of $i$-th sample and $c$ is the number of categories. To describe the missing cases, we let $\mathbf{W} \in \{0,1\}^{n\times m}$ be the missing-view indicator matrix, where $\mathbf{W}_{ij} = 1$ represent $j$-th view of sample $i$ is available, otherwise $\mathbf{W}_{ij} = 0$. Similarly, we define $\mathbf{G} \in \{0,1\}^{n\times c}$ as the missing-label indicator matrix, where $\mathbf{G}_{ij} = 1$ represents $j$-th category of sample $i$ is known, otherwise $\mathbf{G}_{ij} = 0$. All missing information of feature $\mathbf{X}$ and label $\mathbf{Y}$ will be set as '0' in the data-preparation stage. And our goal of multi-view multi-label learning is to train a model which can correctly predict multiple categories for each input sample.

## Related Works

A matrix completion based *MvMlC* method, termed as lrMMC, which forces common subspace to be low rank to satisfy the assumption of matrix completion (Liu et al. 2015). However, lrMMC is incapable of adapting to missing-view or missing-label data. MVL-IV, an incomplete multi-view learning method, attempts to exploit the connections between views (Xu, Tao, and Xu 2015). As another incomplete multi-view single-label learning model, iMSF cleverly divides the incomplete multi-view classification tasks into multiple complete subtasks (Yuan et al. 2012). There is a limitation to both methods that MVL-IV and iMSF only consider the incompleteness of views. On the contrary, MvEL, a method aiming to capture the semantic context and the neighborhood consistency, could only handle the incomplete multi-label case (Zhang et al. 2013). In recent years, there are few approaches to consider double missing cases except iMvWL(Tan et al. 2018) and NAIML(Li and Chen 2021). iMvWL simultaneously maps multi-view features and muti-label information to a common subspace. And a correlation

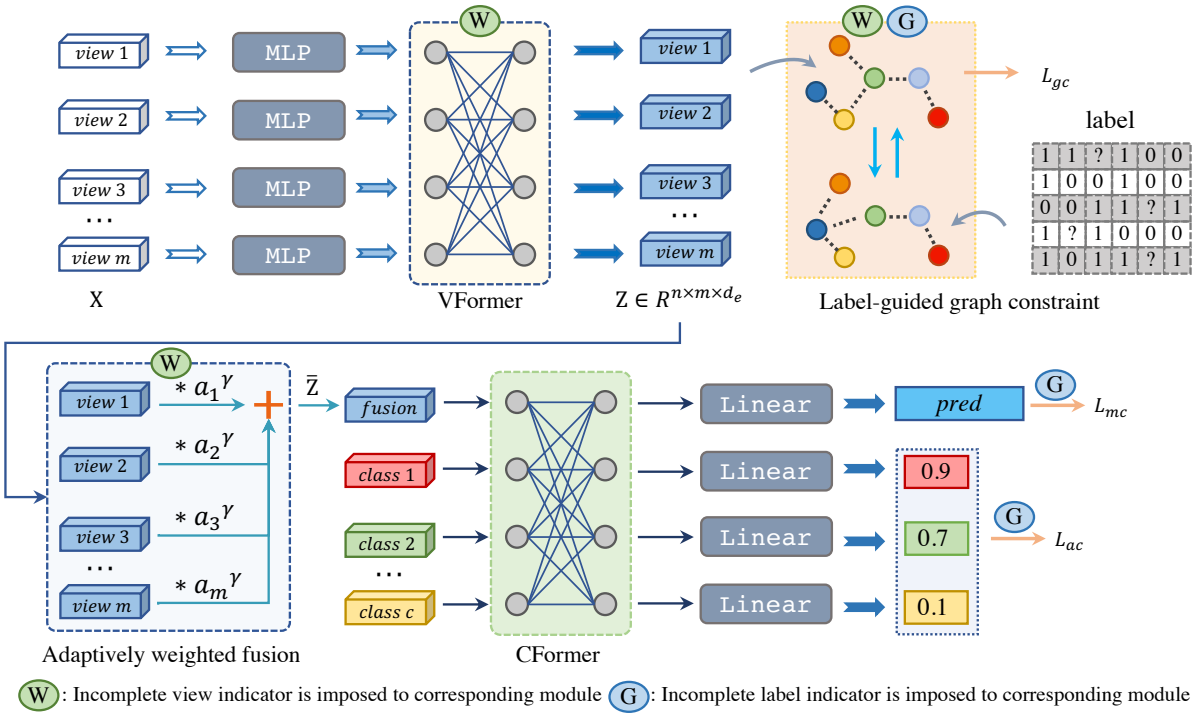



Figure 1: The main structure of our LMVCAT. The MLP layer is composed of two linear layers, GELU activation function, and two dropout layers. And *class* 1 to *class* c denote the $c$ category tokens. Incomplete view indicator and incomplete label indicator are imported to corresponding modules and losses.

matrix is introduced to enhance the projection from label space to embedding subspace. NAIML copes with the double incomplete problem based on the low-rank hypothesis of sub-label matrix, which also implicitly exploits the sub-class correlations. In addition, because the iMvWL and NAIML are well suited to datasets with both view and label incompleteness, we focus on evaluating these two methods in our comparison experiments.

## Method

In this section, we elaborate on each component of our model, including view-aware transformer, label-guided graph constraint, multi-view adaptively weighted fusion, and category-aware transformer.

### VFormer

As we all know, the key to success of multi-view learning is the complementarity among views, which are not available in single-view data. To this end, we design a view-aware transformer encoder (**VFormer** for short) to aggregate complementary information during the cross-view interaction. Before this, it needs to be considered that different views may have different feature dimensions, so for convenience, we map the original multi-view data to the embedding space with the same dimension by a stack of Multilayer Perceptrons (MLP) $\{\boldsymbol{\Phi}_\theta^{(v)}\}_{v=1}^{m}$, which can also be seen as a preliminary feature extraction operation, *i.e.*, $\boldsymbol{\Phi}_\theta^{(v)} : \mathbf{X}^{(v)} \in \mathbb{R}^{n\times d_v} \rightarrow \widehat{\mathbf{X}}^{(v)} \in \mathbb{R}^{n\times d_e}$. These embedding features of multiple views are stacked into a feature sequence $\widehat{\mathbf{X}} \in \mathbb{R}^{n\times m\times d_e}$, like a sentence embedding tensor composed of some word embedding vectors, as the input tensor of the VFormer. The structure of our VFormer is similar to that of the encoder in the typical transformer (Vaswani et al. 2017), and the main difference is that we introduce a missing view indicator matrix in the calculation of multi-head self-attention scores to prevent missing views from participating in the calculation of attention scores. Our masked multi-head self-attention encoder is characterized as follows:

For each sample embedding $\widehat{\mathbf{X}}_i \in \mathbb{R}^{m\times d_e}$, we project it linearly to get its queries, keys, and values by $h$ groups projective matrices, *i.e.*, $\{\mathbf{W^q}_t, \mathbf{W^k}_t, \mathbf{W^v}_t\}_{t=1}^{h}$ with head number $h$. To mask the embedding features according to missing-view distribution, we define a *fill* function to fill zero value with $-1e^9$ and construct a mask matrix of sample $i$: $\mathbf{M}_i = w_i^T w_i \in \mathbb{R}^{m\times m}$, where $w_i$ is $i$-th row vector of $\mathbf{W}$. Then we calculate view correlations $\mathbf{A}_t$ and output $\mathbf{H}_t$:

$$\begin{aligned} \mathbf{A}_t &= softmax\Big(fill\big((\widehat{\mathbf{X}}_i\mathbf{W^q}_t)(\widehat{\mathbf{X}}_i\mathbf{W^k}_t)^T\mathbf{M}_i\big)/\sqrt{d_h}\Big) \\ \mathbf{H}_t &= \mathbf{A}_t(\widehat{\mathbf{X}}_i\mathbf{W^v}_t), \end{aligned} \tag{1}$$

where $d_h = d_e/h$, $\mathbf{W^q}_t, \mathbf{W^k}_t, \mathbf{W^v}_t \in \mathbb{R}^{d_e\times d_h}$, and $\mathbf{H}_t \in \mathbb{R}^{m\times d_h}$. The masked self-attention mechanism is shown in the Fig. 2, it is worth noting that we fill the attention values with $-1e^9$ so that the softmax will ignore the corresponding

missing views when calculating the attention scores. And then, we concatenate all outputs to produce a new embedding feature *w.r.t* sample $i$: $\mathbf{H} = Concat(\mathbf{H}_1, ..., \mathbf{H}_t) \in \mathbb{R}^{m \times d_e}$. To sum up, all views of the same sample will exchange information during the parallel encoding process in our VFormer. As a result, the private information of each view is shared to some extent by other views. Other operations on VFormer are shown in the Fig. 3(a). Finally, our VFormer can be formulated as: $\boldsymbol{\Gamma} : \widehat{\mathbf{X}} \to \mathbf{Z} \in \mathbb{R}^{n \times m \times d_e}$.

## Label-Guided Graph Constraint

Unlike single-label samples maintaining an invariant label distance ($\sqrt{2}$ by Euclidean distance), multi-label samples naturally hold uneven distribution in label space, which offers the possibility to guide the high-level representation learning based on label similarity. Simply put, the label manifold assumption means that if two samples are similar, their labels should also be similar (Wu et al. 2014). In turn, we utilize label similarity to construct a sample-level graph constraint to guide the representation learning. As shown in Fig. 3(b), the similarity vector from sample 1 to other samples is calculated to guide the embedding encoding of sample 1. Before that, we define the label similarity matrix $\mathbf{T}$:

$$\mathbf{T} = \big(\mathbf{Y} \odot \mathbf{G}\big)\big(\mathbf{Y} \odot \mathbf{G}\big)^T ./ \big(\mathbf{G}\mathbf{G}^T\big), \tag{2}$$

where '$\odot$' is Hadamard product and '$./$' denotes the division of corresponding elements. Notably, $\mathbf{T} \in [0, 1]^{n \times n}$ is normalized by $\mathbf{G}\mathbf{G}^T$, where $(\mathbf{G}\mathbf{G}^T)_{ij}$ is referring to the number of known categories in both $\mathbf{Y}_i$ and $\mathbf{Y}_j$. In other words, for two samples, the larger the number of common positive tags, the more similar they are. In addition, the similarity of two embedding features is calculated in cosine space, *i.e.*, for sample $i$ and $j$, their similarity in view $v$ is defined as:

$$\mathbf{S}_{ij}^{(v)} = \Big(\frac{\langle z_i^{(v)} \cdot z_j^{(v)} \rangle}{\|z_i^{(v)}\| \|z_j^{(v)}\|} + 1\Big)/2, \tag{3}$$

where $\langle \cdot \rangle$ denotes the vector dot product operation. $z_i^{(v)}$ and $z_j^{(v)}$ are two embedding features from view $v$ that are output by VFormer. In order to learn sample neighbor relationships from labels, we let label similarity be the target and feature similarity be the learning object. The graph constraint loss $\boldsymbol{\mathcal{L}}_{gc}$ is formulated as:

$$\begin{aligned} \boldsymbol{\mathcal{L}}_{gc} = -\frac{1}{2mN} \sum_{v=1}^{m} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \big(\mathbf{T}_{ij} \log \mathbf{S}_{ij}^{(v)} \\ + \big(1 - \mathbf{T}_{ij}\big) \log \big(1 - \mathbf{S}_{ij}^{(v)}\big)\big)\big(\mathbf{W}_{iv}\mathbf{W}_{jv}\big), \end{aligned} \tag{4}$$

where $N = \sum_{i,j} \mathbf{W}_{iv}\mathbf{W}_{jv}$ denotes the number of available sample pairs in view $v$. We introduce $\mathbf{W}_{iv}\mathbf{W}_{jv}$ to mask the calculation of loss *w.r.t* missing views.

## Adaptively Weighted Fusion

As mentioned above, VFormer encodes an embedding feature for each view of each sample (*i.e.*, $\mathbf{Z} \in \mathbb{R}^{n \times m \times d_e}$). In order to obtain a consistent common representation to uniquely describe the corresponding sample, we propose an adaptively weighted fusion strategy to fuse multi-view embedding features before final classification. The fusion feature $\bar{z}_i$ of sample $i$ is defined as follows:

$$\bar{z}_i = \sum_{v=1}^{m} \frac{e^{a_v^\gamma} z_i^{(v)} \mathbf{W}_{iv}}{\sum_v e^{a_v^\gamma} \mathbf{W}_{iv}}, \tag{5}$$

where $a_v$ is a learnable scalar weight of $v$-th view and $\gamma$ is a adjustment factor. Apparently simple as Eq. (5) seems, it serves two purposes: i) Distinct from other methods to treat all views equally, we flexibly assign each view different weighting coefficients, which help to maintain or highlight the differences of discriminative ability among views. ii) Unusable views must be ignored in multi-view fusion to avoid negative effects, so missing-view indicator matrix is introduced in our fusion module. Stack all $\bar{z}_i$ and we can get output tensor $\bar{\mathbf{Z}} \in \mathbb{R}^{n \times d_e}$ for next classification.

## CFormer and Classifier

In the real world, multiple categories of samples are not independent of each other. How to leverage the multi-label correlations to make our model more discriminative is the

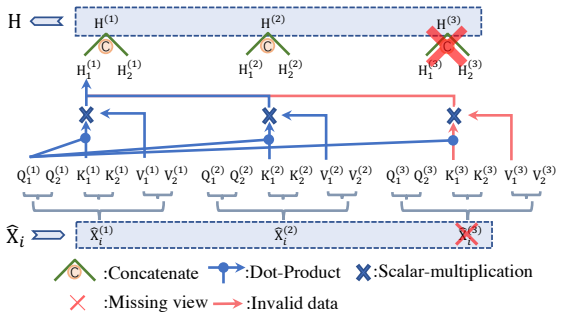



Figure 2: The masked multi-head self-attention mechanism of VFormer taking two heads as an example.

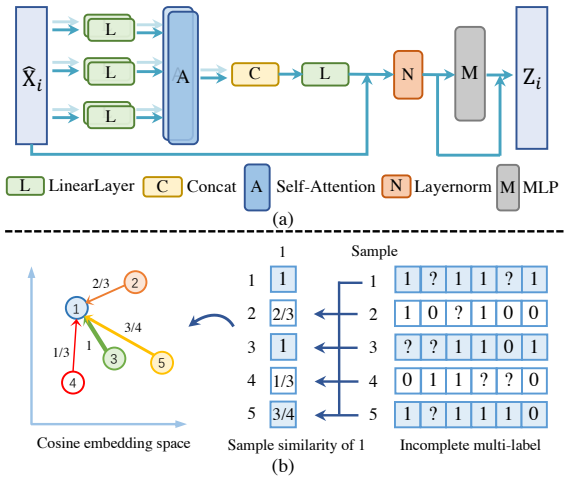



Figure 3: (a) The brief construction of VFormer taking sample $i$ as an example; (b) The simple schematic diagram of label-guided graph constraint.

main concern in the subsection. Instead of learning a category correlation graph, similar to (Lanchantin et al. 2021), we directly map each category to the feature space and leverage the self-attention mechanism to capture the category correlations. In more detail, $c$ class tokens $\left\{cls^i \in \mathbb{R}^{d_e}\right\}_{i=1}^{c}$ are randomly initialized before training and input to category-aware transformer (**CFormer** for short) with fusion features of samples (*i.e.*, the input of CFormer is $\{\bar{z}_i, cls^1, ..., cls^c\}_{i=1}^{n}$). Similar to the structure of VFormer, CFormer allows the fusion features and category tokens to share information, which enjoys two major benefits: On the one hand, the view-fusion features aggregate all subcategory information according to similarities among them, which makes the embedding features encoded by CFormer closer to their relevant class tokens. On the other hand, the information interaction across categories implicitly promotes the learning of category relevance. It should be pointed out that we do not introduce the missing-label mask here because mining the category correlation information requires the participation of all class tokens. And for any sample $i$, the output tensor of our CFormer is $\{\widehat{z}_i, cls_i^1, ..., cls_i^c\} \in \mathbb{R}^{(c+1)\times d_e}$. We split the output into two parts, *i.e.*, consensus representation $\{\widehat{z}_i\}$ for the main classification purpose and $\{cls_i^1, ..., cls_i^c\}$ for the representation learning of class tokens.

As shown in Fig. 1, $c+1$ linear classifiers $\{\boldsymbol{\Psi}_z, \{\boldsymbol{\Psi}_i\}_{i=1}^{c}\}$ are connected in parallel to the outputs of CFormer, where the classifier $\boldsymbol{\Psi}_z$ predicts the final result $\mathbf{P}_z \in [0,1]^{n\times c}$. While each other classifier $\boldsymbol{\Psi}_i$ only predicts the value of its own category so we can get another prediction $\mathbf{P}_c \in [0,1]^{n\times c}$ from all class tokens. Finally, we define a masked binary cross-entropy function as our multi-label classification loss:

$$\begin{aligned}\boldsymbol{\mathcal{L}}_{mbce} = -\frac{1}{C}\sum_{i=1}^{n}\sum_{j=1}^{c}\Big(&\mathbf{Y}_{ij}\log(\mathbf{P}_{ij}) \\ &+ (1-\mathbf{Y}_{ij})\log(1-\mathbf{P}_{ij})\Big)\mathbf{G}_{ij},\end{aligned} \tag{6}$$

where $C = \sum_{i,j}\mathbf{G}_{\boldsymbol{i,j}}$ denotes the number of available labels and $\mathbf{P}$ is the prediction. $\mathbf{G}$ is introduced to prevent unknown labels from participating in the calculation of loss. As a result, we can obtain a main classification loss $\boldsymbol{\mathcal{L}}_{mc}$ and an ancillary classification loss $\boldsymbol{\mathcal{L}}_{ac}$ according to $\boldsymbol{\mathcal{L}}_{mbce}$ for $\mathbf{P}_z$ and $\mathbf{P}_c$, respectively. Overall, our total loss function is :

$$\boldsymbol{\mathcal{L}} = \boldsymbol{\mathcal{L}}_{mc} + \alpha\boldsymbol{\mathcal{L}}_{gc} + \beta\boldsymbol{\mathcal{L}}_{ac}, \tag{7}$$

where $\alpha$ and $\beta$ are penalty coefficients.

# Experiments

## Experimental Setting

**Datasets.** In the experiments, we use five multi-view multi-label datasets as same as (Tan et al. 2018; Guillaumin, Verbeek, and Schmid 2010; Li and Chen 2021): (1) *corel5k* (Duygulu et al. 2002): The corel5k dataset contains 5000 image samples with 260 classes and we use 4999 samples in the experiments. (2) *Pascal07* (Everingham et al. 2009): The popular Pascal07 dataset has 9963 images covering 20 types of tags. (3) *Espgame* (Von Ahn and Dabbish 2004): 20770 samples and 268 categories of Espgame are used in our experiments. It is no doubt that it is a large-scale database. (4) *IAPRTC12* (Grubinger et al. 2006): As a benchmark database, IAPRTC12 is composed of 20,000 high-quality natural images and we use 19627 images and 291 tags in the experiments. (5) *Mirflickr* (Huiskes and Lew 2008): The Mirflickr is collected from the social photography site Flickr, which is made up by 25000 images. 38 types of annotations are selected in the experiments. For the above five multi-view multi-label datasets, each dataset contains six views, *i.e.*, GIST, HSV, HUE, SIFT, RGB, and LAB.

**Data processing.** To evaluate the performance of various multi-view multi-label learning methods on incomplete datasets, following (Tan et al. 2018), we treat the five datasets as follows to simulate the incomplete case: (1) For each view, we randomly remove $50\%$ of samples while guaranteeing that at least one view is available for each sample. (2) For each category, we randomly select $50\%$ of positive tags and negative tags as unknown labels.

**Comparison.** We select eight top methods to compare to our LMVCAT. Six methods, *i.e.*, lrMMC, MVL-IV, MvEL, iMSF, iMvWL, and NAIML, are introduced in section . In addition, we add two advanced methods, named C2AE (Yeh et al. 2017) and GLOCAL (Zhu, Kwok, and Zhou 2018), to expand the evaluation experiments. C2AE is a canonical correlated autoencoder network, which learns a latent embedding space to bridge feature representations and label information. GLOCAL exploits the global and local label correlations on the representation learning, which is also an application of label manifold. However, C2AE and GLOCAL are both single-view multi-label classification methods, and only iMvWL and NAIML are designed for the datasets with both missing views and missing labels. Therefore, we have to do extra alterations to the rest of the methods. Like (Tan et al. 2018) and (Li and Chen 2021), average values of available views are filled into unusable views for MvEL and lrMMC. And as to MVL-IV and iMSF, we set missing tags to be negative tags. C2AE and GLOCAL are independently conducted on each view and the best results are reported. All parameters of these comparison methods are set as recommended in their papers or codes for a fair comparison.

**Evaluation.** Different from (Tan et al. 2018) and (Li and Chen 2021), we only select AP (Average Precision), RL (Ranking Loss), and AUC (adapted Area Under Curve) as our metrics due to the weak discrimination of HL (Hamming Loss). Note that 1-RL is used instead of RL in our results so that the higher the value, the better the performance.

The implementation of our method is based on MindSpore and Pytorch framework.

## Experimental Results and Analysis

Table 1 to 3 show the performance of the nine methods on five incomplete datasets (some of the results are quoted from (Li and Chen 2021) and (Tan et al. 2018)). Table 4 lists the results on the datasets with complete views and labels. All experiments are repeated 10 times to avoid outlier results as much as possible, and we list the mean and standard deviation of 10 tests in these tables. Values in parentheses represent the standard deviation. Fig. 4 and Fig. 5 show more re-

| Data | lrMMC | MVL-IV | MvEL | iMSF | C2AE | GLOCAL | iMvWL | NAIML | ours |
|---|---|---|---|---|---|---|---|---|---|
| Corel5k | .240(.002) | .240(.001) | .204(.002) | .189(.002) | .227(.008) | 0.285(0.004) | .283(.007) | .309(.004) | **.382(.004)** |
| Pascal07 | .425(.003) | .433(.002) | .358(.003) | .325(.000) | .485(.008) | 0.496(0.004) | .441(.017) | .488(.003) | **.519(.005)** |
| Espgame | .188(.000) | .189(.000) | .132(.000) | .108(.000) | .202(.006) | 0.221(0.002) | .242(.003) | .246(.002) | **.294(.004)** |
| Iaprtc12 | .197(.000) | .198(.000) | .141(.000) | .101(.000) | .224(.007) | 0.256(0.002) | .235(.004) | .261(.001) | **.317(.003)** |
| Mirflickr | .441(.001) | .449(.001) | .375(.000) | .323(.000) | .505(.008) | 0.537(0.002) | .495(.012) | .551(.002) | **.594(.005)** |

Table 1: The AP values of nine methods on the five datasets with 50% missing-view ratio, 70% training samples, and 50% missing-label ratio. The best resluts are marked in bold.

| Data | lrMMC | MVL-IV | MvEL | iMSF | C2AE | GLOCAL | iMvWL | NAIML | ours |
|---|---|---|---|---|---|---|---|---|---|
| Corel5k | .762(.002) | .756(.001) | .638(.003) | .709(.005) | .804(.010) | 0.840(0.003) | .865(.003) | .878(.002) | **.880(.002)** |
| Pascal07 | .698(.003) | .702(.001) | .643(.004) | .568(.000) | .745(.009) | 0.767(0.004) | .737(.009) | .783(.001) | **.811(.004)** |
| Espgame | .777(.001) | .778(.000) | .683(.002) | .722(.002) | .772(.006) | 0.780(0.004) | .807(.001) | .818(.002) | **.828(.002)** |
| Iaprtc12 | .801(.000) | .799(.001) | .725(.001) | .631(.000) | .806(.005) | 0.825(0.002) | .833(.003) | .848(.001) | **.870(.001)** |
| Mirflickr | .805(.000) | .804(.001) | .746(.001) | .665(.001) | .821(.003) | 0.832(0.001) | .836(.002) | .850(.001) | **.865(.003)** |

Table 2: The 1-RL values of nine methods on the five datasets with 50% missing-view ratio, 70% training samples, and 50% missing-label ratio. The best resluts are marked in bold.

| Data | lrMMC | MVL-IV | MvEL | iMSF | C2AE | GLOCAL | iMvWL | NAIML | ours |
|---|---|---|---|---|---|---|---|---|---|
| Corel5k | .763(.002) | .762(.001) | .715(.001) | .663(.005) | .806(.010) | 0.843(0.003) | .868(.003) | .881(.002) | **.883(.002)** |
| Pascal07 | .728(.002) | .730(.001) | .686(.005) | .620(.001) | .765(.010) | 0.786(0.003) | .767(.012) | .811(.001) | **.834(.004)** |
| Espgame | .783(.001) | .784(.000) | .734(.001) | .674(.003) | .777(.006) | 0.784(0.004) | .813(.002) | .824(.002) | **.833(.002)** |
| Iaprtc12 | .805(.000) | .804(.001) | .746(.001) | .665(.001) | .807(.005) | 0.830(0.001) | .836(.002) | .850(.001) | **.872(.001)** |
| Mirflickr | .806(.001) | .807(.000) | .761(.000) | .715(.001) | .810(.004) | 0.828(0.001) | .794(.015) | .837(.001) | **.853(.003)** |

Table 3: The AUC values of nine methods on the five datasets with 50% missing-view ratio, 70% training samples, and 50% missing-label ratio. The best resluts are marked in bold.


(a) different missing-view ratios (b) different missing-label ratios

Figure 4: The results on Corel5k dataset with (a) different missing-view ratios and a 50% missing-label ratio and (b) a 50% missing-view ratio and different missing-label ratios.


(a) Corel5k (b) Mirflickr

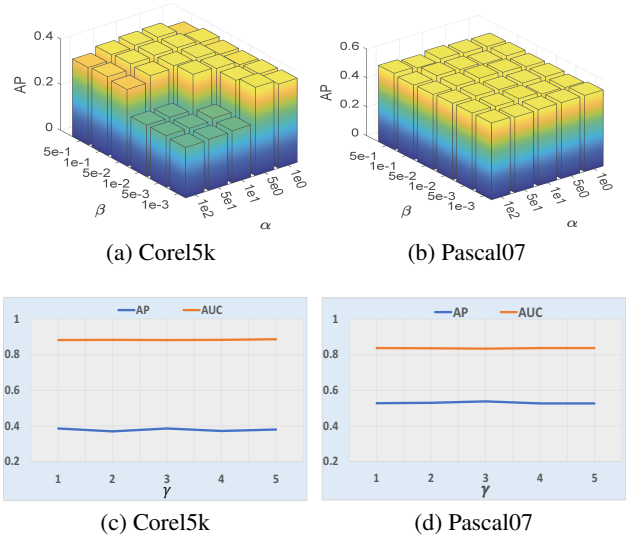
Figure 5: The results on (a) Corel5k dataset and (b) Mirflickr dataset with a 50% missing-view ratio, a 50% missing-label ratio, and different training sample ratios.

sults *w.r.t* different missing and training sample rates. From Table 1 to 4, we can summarize the following observations:

- Our method achieves overwhelming advantages on all three metrics of the five datasets, especially on the most representative AP metric. For example, the AP value of our model exceeds the second-best NAIML by about 7% and 5% on the Corel5k and Espgame datasets respectively, which verifies the superiority of our method.
- Both iMvWL and NAIML reach relatively better performance compared to the first four methods, which benefits from good compatibility to double incomplete data. As a deep method with stronger fitting ability, C2AE performs mediocrely due to the lack of multi-view complementary information. Though GLOCAL is a single-view method, the full exploitation of label correlations helps it achieve surprising results on several datasets.
- From Table 4, we can find that, although our LMVCAT is an incomplete method, it still works well with complete datasets. Specifically, on the Corel5k dataset, the AP value of our method is about 14 percentage points

| Dateset | Metric | C2AE | GLOCAL | iMvWL | NAIML | ours |
|---|---|---|---|---|---|---|
| Corel5k | AP | .353 | .386 | .313 | .327 | **.521** |
| | 1-RL | .870 | .891 | .884 | .890 | **.928** |
| | AUC | .873 | .895 | .887 | .893 | **.930** |
| Pascal07 | AP | .584 | .610 | .468 | .496 | **.629** |
| | 1-RL | .835 | .866 | .763 | .795 | **.878** |
| | AUC | .851 | .879 | .793 | .822 | **.892** |
| Espgame | AP | .269 | .264 | .260 | .251 | **.385** |
| | 1-RL | .832 | .804 | .817 | .825 | **.876** |
| | AUC | .837 | .810 | .822 | .830 | **.880** |
| Iaprtc12 | AP | .316 | .330 | .250 | .267 | **.436** |
| | 1-RL | .868 | .871 | .842 | .825 | **.918** |
| | AUC | .869 | .877 | .843 | .830 | **.918** |
| Mirflickr | AP | .593 | .643 | .493 | .555 | **.684** |
| | 1-RL | .854 | .876 | .806 | .847 | **.905** |
| | AUC | .845 | .869 | .789 | .839 | **.889** |

Table 4: The experimental results on the five datasets with full views, full labels and 70% training samples. The best results are marked in bold.




(a) Corel5k (b) Pascal07

(c) Corel5k (d) Pascal07

Figure 6: The results of different parameter selections on (a,c) Corel5k and (b,d) Pascal07 datasets with half the available views and known labels and a 70% training sample rate.

ahead of the second-best GLOCAL.

As shown in Fig. 4, we respectively plot the histogram of three metrics in different missing-view and missing-label ratios by fixing another ratio on the corel5k dataset. These results verify that both incomplete views and partial labels are harmful for efficient classification. In addition, an interesting phenomenon is that when the incompleteness ratio is relatively small, missing labels have a greater impact on the performance, and conversely, the negative impact of missing views is greater. As shown in Fig. 5, we conduct experiments on Corel5k and Mirflickr datasets with 50% available views,

| method | Corel5k | | | Espgame | | |
|---|---|---|---|---|---|---|
| | AP | 1-RL | AUC | AP | 1-RL | AUC |
| $\mathcal{V}$ | .347 | .871 | .874 | .284 | .822 | .827 |
| $\mathcal{V}+\mathcal{A}$ | .348 | .872 | .877 | .286 | .824 | .829 |
| $\mathcal{V}+\mathcal{A}+\mathcal{C}$ | .359 | .879 | .883 | .288 | .828 | .833 |
| $\boldsymbol{\mathcal{V}+\mathcal{A}+\mathcal{C}+\mathcal{G}}$ | **.382** | **.880** | **.883** | **.294** | **.828** | **.833** |
| *w/o* view mask | .352 | .866 | .870 | .274 | .817 | .823 |
| *w/o* label mask | .365 | .880 | .883 | .287 | .827 | .832 |

Table 5: The ablation experiments on two datasets with 50% missing-view ratio, 50% missing-label ratio, and 70% training samples. $\mathcal{V}$ is the backbone with VFormer; $\mathcal{A}$ denotes the adaptively weighted strategy; $\mathcal{C}$ represents the CFormer; and $\mathcal{G}$ means the graph constraint.

50% known tags and different training sample ratios. It is clear that, as the proportion of training samples in the total increases, the performance of our method gradually goes up.

## Hyperparameters Study

There are three hyperparameters, *i.e.*, $\alpha$, $\beta$, and $\gamma$ in our LMVCAT. To study the optimal parameters selection, we list the performance of our method on different parameter combinations in Fig. 6. All datasets used in parameters study are with 50% missing-view ratio, 50% missing-label ratio, and 70% training samples. As can be seen in Fig. 6 (a) and (b), for Corel5k dataset, the optimal parameters $\alpha$ and $\beta$ are located in the range of $[5, 10]$ and $[0.05, 0.5]$, respectively, and for Pascal07 dataset, the optimal parameters $\alpha$ and $\beta$ are located in the range of $[10, 100]$ and $[0.001, 0.5]$, respectively. As to $\gamma$, obviously, our method is insensitive to it from Fig. 6 (c) and (d). In our experiments, $\gamma$ is uniformly set to 2.

## Ablation Study

To study the effectiveness of each component of our method, we perform experiments with following altered methods. For the complete LMVCAT, we remove graph constraint, CFormer, and adaptively weighted module sequentially. Besides, we take off the missing-view indicator and missing-label indicator in our model so that the incomplete views and weak labels are completely exposed to the network. As can be seen in Table 5, our label-guided graph constraint plays a key role and all experiments confirm that each component of our LMVCAT is beneficial and necessary.

## Conclusion

In this paper, we propose a general transformer-based framework for *MvMlC*, which skillfully exploits label manifold to guide the representation learning. And its innovative view- and category-awareness realize multi-view information interaction and multi-category discrimination fusion. An adaptively weighted fusion strategy is also introduced to balance the view-specific contribution. In addition, missing-view and weak label problems are specifically avoided throughout the network. Extensive experiments support the conclusion of the effectiveness of our method.

## Acknowledgments

This work is supported by Shenzhen Science and Technology Program under Grant RCBS20210609103709020, GJHZ20210705141812038, Shenzhen Fundamental Research Fund under Grant GXWD20220811173317002, and CAAI-Huawei MindSpore Open Fund under Grant CAAIXSJLJJ-2022-011C.

## References

Chen, M.-S.; Liu, T.; Wang, C.-D.; Huang, D.; and Lai, J.-H. 2022. Adaptively-weighted Integral Space for Fast Multiview Clustering. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3774–3782.

Chen, M.-S.; Wang, C.-D.; and Lai, J.-H. 2022. Low-rank Tensor Based Proximity Learning for Multi-view Clustering. *IEEE Transactions on Knowledge and Data Engineering*.

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 *European Conference on Computer Vision*, 97–112.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2009. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–308.

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 Conference on Language Resources and Evaluation*, 1–11.

Guillaumin, M.; Verbeek, J.; and Schmid, C. 2010. Multimodal semi-supervised learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 902–909.

Hu, S.; Lou, Z.; and Ye, Y. 2021. View-Wise Versus Cluster-Wise Weight: Which Is Better for Multi-View Clustering? *IEEE Transactions on Image Processing*, 31: 58–71.

Hu, S.; Shi, Z.; and Ye, Y. 2020. DMIB: Dual-correlated multivariate information bottleneck for multiview clustering. *IEEE Transactions on Cybernetics*.

Huang, C.; Liu, C.; Wen, J.; Wu, L.; Xu, Y.; Jiang, Q.; and Wang, Y. 2022a. Weakly Supervised Video Anomaly Detection via Self-Guided Temporal Discriminative Transformer. *IEEE Transactions on Cybernetics*.

Huang, C.; Liu, C.; Zhang, Z.; Wu, Z.; Wen, J.; Jiang, Q.; and Xu, Y. 2022b. Pixel-Level Anomaly Detection via Uncertainty-Aware Prototypical Transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, 521–530.

Huang, J.; Qin, F.; Zheng, X.; Cheng, Z.; Yuan, Z.; Zhang, W.; and Huang, Q. 2019. Improving multi-label classification with missing labels by learning label-specific features. *Information Sciences*, 492: 124–146.

Huiskes, M. J.; and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *ACM International Conference on Multimedia Information Retrieval*, 39–43.

Lanchantin, J.; Wang, T.; Ordonez, V.; and Qi, Y. 2021. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16478–16488.

Li, L.; and He, H. 2020. Bipartite graph based multi-view clustering. *IEEE transactions on knowledge and data engineering*, 34(7): 3111–3125.

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*.

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*, 1–1.

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

Liu, C.; Wu, Z.; Wen, J.; Xu, Y.; and Huang, C. 2022. Localized Sparse Incomplete Multi-view Clustering. *IEEE Transactions on Multimedia*.

Liu, M.; Luo, Y.; Tao, D.; Xu, C.; and Wen, Y. 2015. Low-Rank Multi-View Learning in Matrix Completion for Multi-Label Image Classification. In *AAAI Conference on Artificial Intelligence*, volume 29.

Liu, X.; Zhou, S.; Wang, Y.; Li, M.; Dou, Y.; Zhu, E.; and Yin, J. 2017. Optimal neighborhood kernel clustering with multiple kernels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Lu, X.; Zhu, L.; Cheng, Z.; Nie, L.; and Zhang, H. 2019. Online multi-modal hashing with dynamic query-adaption. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, 715–724.

Luo, X.; Pu, Z.; Xu, Y.; Wong, W. K.; Su, J.; Dou, X.; Ye, B.; Hu, J.; and Mou, L. 2021. MVDRNet: Multi-view diabetic retinopathy detection by combining DCNNs and attention mechanisms. *Pattern Recognition*, 120: 108104.

Luo, X.; Wang, W.; Xu, Y.; Lai, Z.; Jin, X.; Zhang, B.; and Zhang, D. 2023. A deep convolutional neural network for diabetic retinopathy detection via mining local and long-range dependence. *CAAI Transactions on Intelligence Technology*.

Sun, S.; and Zong, D. 2020. Lcbm: a multi-view probabilistic model for multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8): 2682–2696.

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

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Neural Information Processing Systems*, 30.

Von Ahn, L.; and Dabbish, L. 2004. Labeling images with a computer game. In *SIGCHI Conference on Human Factors in Computing Systems*, 319–326.

Wang, L.; Gong, Y.; Ma, X.; Wang, Q.; Zhou, K.; and Chen, L. 2022. IS-MVSNet: Importance Sampling-Based MVS-Net. In *European Conference on Computer Vision*, 668–683. Springer.

Wen, J.; Zhang, Z.; Fei, L.; Zhang, B.; Xu, Y.; Zhang, Z.; and Li, J. 2022. A survey on incomplete multiview clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.

Wen, J.; Zhang, Z.; Xu, Y.; Zhang, B.; Fei, L.; and Liu, H. 2019. Unified embedding alignment with missing views inferring for incomplete multi-view clustering. In *AAAI conference on artificial intelligence*, volume 33, 5393–5400.

Wu, B.; Liu, Z.; Wang, S.; Hu, B.-G.; and Ji, Q. 2014. Multi-label learning with missing labels. In *2014 22nd International Conference on Pattern Recognition*, 1964–1968. IEEE.

Xu, C.; Tao, D.; and Xu, C. 2015. Multi-view learning with incomplete views. *IEEE Transactions on Image Processing*, 24(12): 5812–5825.

Xu, N.; Guo, Y.; Zheng, X.; Wang, Q.; and Luo, X. 2018. Partial multi-view subspace clustering. In *ACM International conference on multimedia*, 1794–1801.

Yeh, C.-K.; Wu, W.-C.; Ko, W.-J.; and Wang, Y.-C. F. 2017. Learning Deep Latent Space for Multi-Label Classification. In *AAAI Conference on Artificial Intelligence*, volume 31.

Yuan, L.; Wang, Y.; Thompson, P. M.; Narayan, V. A.; Ye, J.; Initiative, A. D. N.; et al. 2012. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *NeuroImage*, 61(3): 622–632.

Zhan, K.; Zhang, C.; Guan, J.; and Wang, J. 2017. Graph learning for multiview clustering. *IEEE transactions on cybernetics*, 48(10): 2887–2895.

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, W.; Zhang, K.; Gu, P.; and Xue, X. 2013. Multi-view embedding learning for incompletely labeled data. In *International Joint Conference on Artificial Intelligence*, 1910–1916.

Zhao, X.; Chen, Y.; Liu, S.; and Tang, B. 2022. Shared-Private Memory Networks for Multimodal Sentiment Analysis. *IEEE Transactions on Affective Computing*.

Zhu, Y.; Kwok, J. T.; and Zhou, Z.-H. 2018. Multi-Label Learning with Global and Local Label Correlation. *IEEE Transactions on Knowledge and Data Engineering*, 30(6): 1081–1094.