Attention-Induced Embedding Imputation for Incomplete Multi-View Partial Multi-Label Classification

Chengliang Liu[1], Jinlong Jia[1], Jie Wen[1*], Yabo Liu[1], Xiaoling Luo[2], Chao Huang[3], Yong Xu[1*]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
[2]College of Computer Science and Software Engineering, Shenzhen University
[3]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University
liucl1996@163.com, jjlong_ls@163.com, jiewen_pr@126.com, yaboliu.ug@gmail.com, xiaolingluoo@outlook.com, huangchao_08@126.com, laterfall@hit.edu.cn

Abstract

As a combination of emerging multi-view learning methods and traditional multi-label classification tasks, multi-view multi-label classification has shown broad application prospects. The diverse semantic information contained in heterogeneous data effectively enables the further development of multi-label classification. However, the widespread incompleteness problem on multi-view features and labels greatly hinders the practical application of multi-view multi-label classification. Therefore, in this paper, we propose an attention-induced missing instances imputation technique to enhance the generalization ability of the model. Different from existing incomplete multi-view completion methods, we attempt to approximate the latent features of missing instances in embedding space according to cross-view joint attention, instead of recovering missing views in kernel space or original feature space. Accordingly, multi-view completed features are dynamically weighted by the confidence derived from joint attention in the late fusion phase. In addition, we propose a multi-view multi-label classification framework based on label-semantic feature learning, utilizing the statistical weak label correlation matrix and graph attention network to guide the learning process of label-specific features. Finally, our model is compatible with missing multi-view and partial multi-label data simultaneously and extensive experiments on five datasets confirm the advancement and effectiveness of our embedding imputation method and multi-view multi-label classification model.

Introduction

Multiple photos taken from different perspectives of the observed object provide a more comprehensive perception, surpassing the limited viewpoint offered by a single perspective (Xu et al. 2023; Fang et al. 2022; Liu et al. 2023e). Combining data from different media can help compensate for the limitations and incompleteness found in individual forms of data (Li and He 2020; Xu et al. 2019). With the explosive growth of multi-view data, multi-view learning demonstrates immense potential in empowering various traditional tasks (Xu et al. 2022a,b; Li, Wan, and He 2021; Liu et al. 2023f). For instance, traditional multi-label classification solely relies on features extracted from a single perception, leading to noticeable limitation in feature awareness. Incorporating multi-view learning into multi-label classification can effectively mitigate this limitation. As a result, a composite multi-view multi-label classification (MvMLC) task has emerged. In recent years, many MvMLC models based on traditional machine learning or deep neural network (DNN) have demonstrated promising performance.

However, in real-world scenarios, it is common to encounter incomplete multi-view data, which poses a great challenge for the design of MvMLC models (Liu et al. 2023b). For example, subtitles and audio are often missing in many ancient video archives. Likewise, multi-label information may also be partially available in various scenarios due to the costliness and vulnerability of manual annotation. And we name the complex composite task "incomplete multi-view partial multi-label classification (iMvPMLC)" in our paper. Besides, according to research conducted by Liu et al., it has been observed to some extent that incomplete views have a greater negative impact on the performance of MvMLC than the incompleteness of label information at the same missing ratio (Liu et al. 2023a). Therefore, we will pay more attention to the missing-view imputation in this paper.

Many approaches have been proposed to tackle the incompleteness issue. Some researchers tried to impute the missing instances in the kernel space by the complete kernel sub-matrix (Trivedi et al. 2010; Liu et al. 2019), or combined missing data imputation and downstream tasks to a unified framework (Liu et al. 2020, 2018). Another route to handle the incomplete views is to introduce prior missing information to skip the unknown views, utilizing only the available data for the classification task, which has also shown promising results (Liu et al. 2023d,a). However, compared with 'skip' or imputation in kernel space, direct recovery of missing data has always been a challenging task with high modeling difficulty and unlimited potential (Liu et al. 2023c). Furthermore, considering that the high-level features extracted by DNN have stronger semantic information (Luo et al. 2023; Fang et al. 2023a), we propose to complete the missing instances in the embedding space, which also avoids the redundancy and noise prevalent in the original data.

To sum up, we propose a novel model called **A**ttention-

*Corresponding authors.

Induced I**M**putation Network (**AIMNet**) for the iMvPMLC task with novel embedding feature imputation. Specifically, we design a two branch multi-view multi-label classification framework with a multi-view feature extraction module and a graph attention network (Veličković et al. 2018) (GAT) based label semantic feature extraction module. For the missing instances in the embedding space, inspired by the widely used attention mechanism (Zhao et al. 2022b; Fang et al. 2023b), we propose an attention-induced feature imputation technique, which approximates missing instances based on available information and instance-pair attention scores. To achieve this, we compute the joint attention matrix by cross-view peak aggregation on multiple view-specific attention matrices. Finally, we incorporate a confidence-based dynamic multi-view late fusion mechanism to further improve the reliability of multi-view fusion. In summary, our paper makes several notable contributions:

- We propose the AIMNet, a novel framework for iMvPMLC, which is able to extract the label semantic feature and instance embedding feature simultaneously. And the AIMNet can effectively handle the learning task with both incomplete views and labels.
- To our knowledge, we are the first to develop a self-attention mechanism to fill missing instances in embedding spaces for iMvPMLC task. And our embedding imputation technique also has great potential to be applied in other deep incomplete multi-view learning methods.
- We perform extensive experiments to confirm that our AIMNet can outperform existing advanced methods on double-missing multi-view multi-label data, while still showing leading performance on complete datasets.

Related Works and Problem Definition

Multi-label Classification

Multi-label classification has been an active research area in the field of machine learning, attracting significant attention and interest. In traditional methods, the multi-label classification task is typically divided into multiple binary classification tasks with label correlations (Read et al. 2009). Building upon this strategy, some studies employ local or global label correlation to model dependency of categories. Huang et al. proposed a method called 'Learning Label Specific Features' (LLSF), which involves extracting label-specific features and establishing correlations among them (Huang et al. 2015). Besides, LLSF reduces the dimensionality of label space, aiding in alleviating computational burden. Furthermore, Ma et al. put forward a network named LDGN that incorporates label-specific semantic components and dual Graph Convolution Network (GCN), achieving impressive results (Ma et al. 2021). However, this method heavily relies on a pre-trained label semantic feature extraction network and is not suitable for cases where the label-related semantic features are weak or non-existent such as pure digital labels. Another work with label semantic learning, termed SSGRL, also relies on a pre-trained tag-specific feature extractor. The difference is that it fuses sample features and label features in the shallow layer of the network and feeds them into graph neural network (GNN) to learn samples' label distribution (Chen et al. 2019).

MvMLC

Applying multi-view learning to multi-label classification task can achieve superior performance than traditional single-view multi-label classification. However, devising an effective multi-view collaborative learning strategy is a challenging task. Zhu et al. proposed a method named Label Space Dimension Reduction (LSDR), which utilizes the Hilbert-Schmidt Independence Criterion (HSIC) technique to reinforce the dependency between different latent spaces with a low complexity (Zhu et al. 2018). Another matrix factorization (MF) based method, called LSA-MML, maximizes interdependencies among latent semantic basis matrices of diverse views in kernel space to learn consistent representation (Zhang et al. 2018). Fang and Zhang developed an innovative method called Consistency and Diversity Multi-View Multi-Label learning (CDMM), which learns independent prediction results for each view and maximizes the dependence between features and labels through HSIC (Fang and Zhang 2012). Considering the incompleteness of multi-view data from both feature and label perspectives brings further challenges to MvMLC. iMvML proposed by Tan et al. first maps incomplete data from multiple views to a shared subspace, and then bridge the shared subspace and the semantic label space based on the projection matrix and learnable label correlations (Tan et al. 2018). Furthermore, assuming that label correlations are locally structured, iMvML imposes a low-rank constraint on the label correlation matrix. Li and Chen proposed another interesting method NAIM3L that focuses on both the global high-rankness and the local low-rankness of the label matrix (Li and Chen 2022). In addition to above traditional methods, Liu et al. proposed a representative deep iMvPMLC framework DICNet that employs stacked autoencoders to extract multi-view features and develops an incomplete instance-level contrastive learning strategy to enhance consensus representation learning (Liu et al. 2023a). Note that all three methods, i.e., iMvML, NAIM3L, and DICNet, choose to handle the missing views or labels by introducing prior index matrix rather than inpainting raw data.

Problem Definition

In order to clearly describe our problem, we define $\left\{\mathbf{X}^{(v)} \in \mathbb{R}^{n \times d_v}\right\}_{v=1}^{m}$ as original multi-view data, in which m, n, and d_v represent the number of view, sample and the original feature dimension of the v-th view, respectively. And $\mathbf{Y} \in \{0,1\}^{n \times c}$ denotes the label index matrix, where c is the number of catagories. For the label vector of i-th sample $\mathbf{Y}_{i,:}$, $\mathbf{Y}_{i,j} = 1$ means that the sample i belongs to the j-th class, otherwise $\mathbf{Y}_{i,j} = 0$. Furthermore, we let $\mathbf{W} \in \{0,1\}^{n \times m}$ and $\mathbf{G} \in \{0,1\}^{n \times c}$ be the missing-view and missing-label index matrix, respectively, where $\mathbf{W}_{i,j}$ will be assigned a value of 1 if the j-th view of i-th sample is available, otherwise $\mathbf{W}_{i,j} = 0$. $\mathbf{G}_{i,j} = 1$ or 0 similarly indicates the certainty of corresponding tag. The missing data in $\left\{\mathbf{X}^{(v)} \in \mathbb{R}^{n \times d_v}\right\}_{v=1}^{m}$ is designated as 'Nan' or ran-

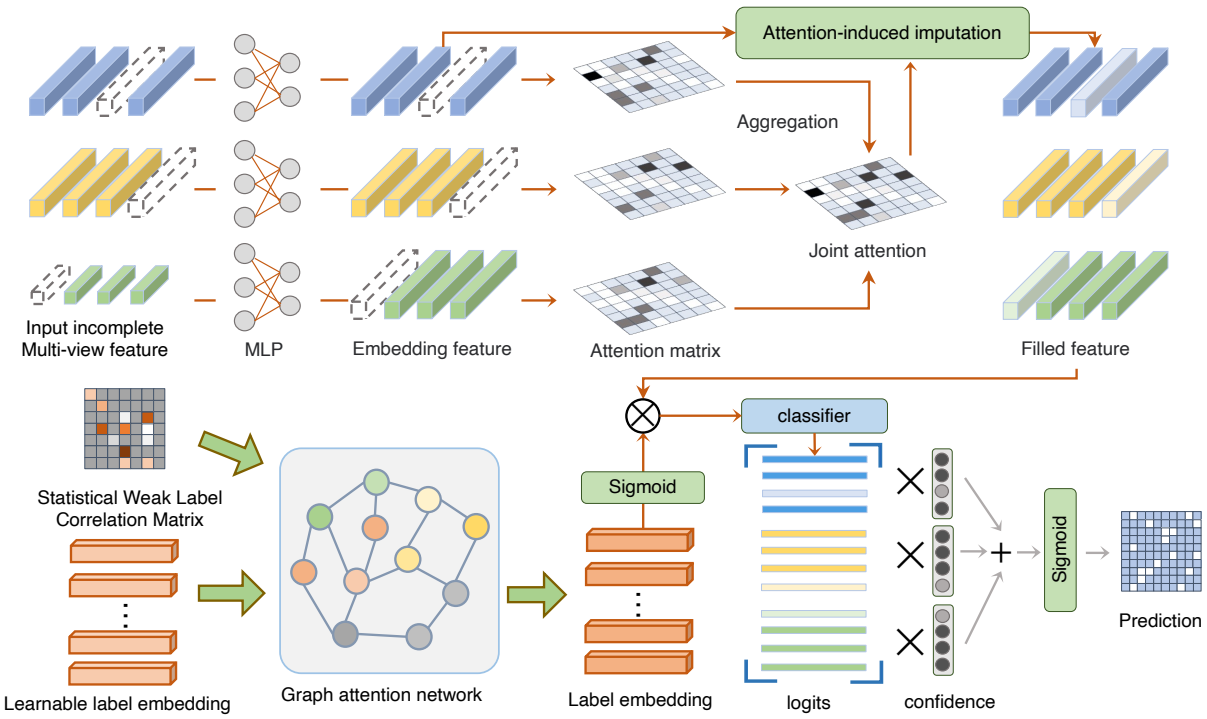

Figure 1: The main flowchart of our AIMNet. Our framework consists of two branches, an instance feature extraction branch and a label semantic extraction branch, which are fused in the embedding space to get the prediction on each view; a multi-view late-fusion method is employed to fuse the predictions of multiple views according to the confidence; missing instances are filled according to the available inter-instance attention.

dom noise to keep the number of instances consistent across views. Similarly, unknown tags are set '0' in $\mathbf{Y}$. The target of our AIMNet is to train a high-precision multi-label classification network on incomplete multi-view partial multi-label data, which can accurately predict the categories of unlabeled incomplete multi-view data.

Method

In this section, we introduce our method in detail from four aspects, namely label semantic representation learning, attention-induced missing view imputation, confidence-based multi-view late fusion, and multi-label classification. Fig .1 is the main flowchart of our AIMNet.

Multi-label Semantic Representation Learning

In the field of multi-label classification, most methods try to establish a direct or indirect mapping from input data to labels, and it is difficult to explicitly model label-specific semantics with correlation (co-occurrence). To address it, in this section, we attempt to learn the semantic representations for each category based on label correlation. We know that label correlation can be represented as a graph structure composed of categories (node) and corresponding paired correlation (edge). A common way to obtain it is by calculating the co-occurrence frequency of any two categories in the training data, i.e., $\mathcal{P}(l_j|l_i)$, which means the probability that j-th category appears when i-th category appears. In view of its low acquisition cost, so in this paper, we construct the label correlation matrix $\mathbf{C}$ by:

$$\mathbf{C}_{i,j} = \mathcal{P}(l_j|l_i) = \frac{\sum_{k=1}^{n} \mathbf{Y}_{k,i}\mathbf{Y}_{k,j}}{\sum_{k=1}^{n} \mathbf{Y}_{k,i}} = \frac{\mathbf{Y}_{:,i}^{T}\mathbf{Y}_{:,j}}{\mathbf{Y}_{:,i}^{T}\mathbf{Y}_{:,i}}. \quad (1)$$

It should be noted that the diagonal elements of $\mathbf{C}$ are set to 0, and $\mathbf{C}$ is asymmetric, that is, the co-occurrence probability $\mathcal{P}(l_j|l_i)$ is not equal to $\mathcal{P}(l_i|l_j)$, and $\mathbf{C}$ without a symmetrical design is more in line with objective laws of real-world data. And then, with the label correlation matrix $\mathbf{C}$, we apply a graph attention networks (GAT) (Veličković et al. 2018) to learn label semantic features. First, we initialize a set of learnable label embedding features $\{h_i\}_{i=1}^{c} \in \mathbb{R}^{c\times c}$, and then feed them into an K-head GAT layer with the label correlation matrix $\mathbf{C}$. For each head, we compute the paired label embedding attention coefficient:

$$\alpha_{ij} = \frac{e^{\sigma_L(a\cdot([Wh_i||Wh_j]))}}{\sum_{k\in\mathcal{N}_i} e^{\sigma_L(a\cdot([Wh_i||Wh_k]))}}, \quad (2)$$

where $||$ is the concatenate operation and σ_L is the LeakyReLU activation function. $a \in \mathbb{R}^{2d_e}$ and $W \in \mathbb{R}^{d_e\times c}$ are the weight parameters. $j \in \mathcal{N}_i$ and $\mathcal{N}_i$ denotes that we only consider the similar classes of class i in the given $\mathbf{C}$, i.e, $\mathbf{C}_{i,j} > 0$. Then, we use the attention coefficients to aggregate the neighbor nodes of each label embedding feature

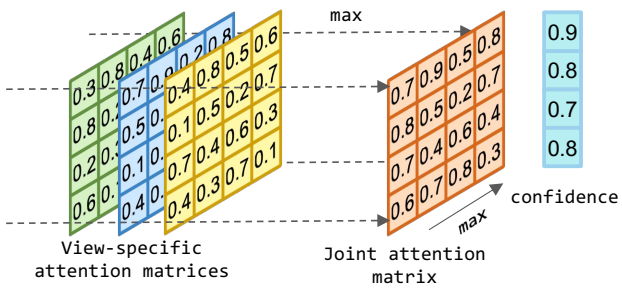

Figure 2: Computation diagram for cross-view joint attention and confidence.

and finally obtain the multi-head output feature:

$$h_i' = \sigma_L\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}_i}\alpha_{ij}^k W^k h_j\right), \tag{3}$$

where α_{ij}^k and W^k are the attention coefficient and weight parameter of k-th head respectively. Finally, we can obtain the output label embedding features $\mathbf{L} = \{h_1', h_2', ..., h_c'\} \in \mathbb{R}^{c\times d_e}$. Overall, in this section, we utilize the graph attention layer to learn multi-label semantic features, which can adaptively aggregate semantic information of relevant labels.

Attention-Induced Missing View Imputation

As shown in Fig. 1, in order to learn the association of label-specific features and samples, like other deep methods, we need to map the raw data to the embedding space to extract effective high-level features. Here we use a set of Multilayer Perceptrons (MLP) to extract multiple view-specific embedding features: $\{\Psi_v : \mathbf{X}^{(v)} \rightarrow \mathbf{Z}^{(v)}\}_{v=1}^m$, where $\mathbf{Z}^{(v)} \in \mathbb{R}^{n\times d_e}$ is the embedding feature of v-th view. Each $\mathbf{Z}^{(v)}$ consists of n instances in view v, however, in the setting of our task, not all n instances are available. To cope with this incompleteness that arises in the embedding space, we propose attention-induced missing view imputation techniques in this section. First, for any view v, we compute the attention score of an instance pair by:

$$\mathbf{A}_{i,j}^{(v)} = e^{f(z_i^{(v)})f(z_j^{(v)T})/\tau}, \tag{4}$$

where $\mathbf{A}^{(v)} \in \mathbb{R}^{n\times n}$ is the attention matrix in terms of instances in view v. τ is the temperature parameter to control the scale and $f(\cdot)$ is the l_2-norm normalization function. Each row of the matrix $\mathbf{A}^{(v)}$ describes the attention of corresponding instance with all other instances, which provides weight support for information aggregation among samples (Vaswani et al. 2017). In our task, rather than augmented information fusion, we focus more on utilizing attention mechanism and available instances to recover or approximately replace unavailable instances, which is the fundamental motivation of our feature imputation method. However, the parts corresponding to missing instances in the attention matrix of each view are not available, so we cannot directly use the view-specific attention matrix to obtain the reconstructed features of the corresponding view. Thus we have to borrow attention information from other views to compensate for the incompleteness of each view itself. Specifically, as shown in Fig. 2, we propose to obtain the cross-view joint attention matrix $\bar{\mathbf{A}} \in \mathbb{R}^{n\times n}$ by maximizing aggregation:

$$\bar{\mathbf{A}}_{i,j} = \max(\mathbf{A}_{i,j}^{(1)}\mathbb{1}_{[\Upsilon_{ij}^1]}, \mathbf{A}_{i,j}^{(2)}\mathbb{1}_{[\Upsilon_{ij}^2]}, ..., \mathbf{A}_{i,j}^{(m)}\mathbb{1}_{[\Upsilon_{ij}^m]}), \tag{5}$$

where $\mathbb{1}_{[\Upsilon_{ij}^v]} = 1$ when condition $\Upsilon_{ij}^v : \{\mathbf{W}_{i,v}\mathbf{W}_{j,v} = 1\}$ is met, i.e., only the attention score between valid instances can be counted. With the joint attention matrix $\bar{\mathbf{A}}$, we can easily get the attention-induced reconstructed features by:

$$\bar{\mathbf{Z}}^{(v)} = f_n(\bar{\mathbf{A}})diag(\mathbf{W}_{:,v})\mathbf{Z}^{(v)}, \tag{6}$$

where $\bar{\mathbf{Z}}^{(v)} \in \mathbb{R}^{n\times d_e}$ is v-th view's reconstructed feature and $diag(\mathbf{W}_{:,v})$ means the diagonal matrix with diagonal $\mathbf{W}_{:,v}$. $f_n(\cdot)$ denotes a normalization operation: $[f_n(\bar{\mathbf{A}})]_{i,j} = \bar{\mathbf{A}}_{i,j}/\sum_{j=1}^n \bar{\mathbf{A}}_{i,j}$. Note that we cannot directly use $\bar{\mathbf{Z}}^{(v)}$ for the subsequent classification process since it is only an approximate representation of missing instances. Therefore, we only select the reconstructed features corresponding to the missing instances in $\bar{\mathbf{Z}}^{(v)}$ to fill in the incomplete original embedding features $\mathbf{Z}^{(v)}$ to obtain the final completed features $\hat{\mathbf{Z}}^{(v)}$. Specifically, for i-th instance in $\hat{\mathbf{Z}}^{(v)}$, we compute it by introducing missing-view index matrix $\mathbf{W}$:

$$\hat{\mathbf{Z}}_{i,:}^{(v)} = \bar{\mathbf{Z}}_{i,:}^{(v)}(1 - \mathbf{W}_{i,v}) + \mathbf{Z}_{i,:}^{(v)}\mathbf{W}_{i,v}. \tag{7}$$

Multi-View Late Fusion and Multi-Label Classification

Up to now, we have label semantic features $\mathbf{L}$ and multi-view embedding features $\{\hat{\mathbf{Z}}^{(v)}\}_{v=1}^m$, which are simultaneously mapped to the space with dimension d_e. And then, we try to bridge them to obtain each sample's classification results corresponding to multiple views. A common concatenation method is to take the dot product of each instance representation with all label semantic features to obtain confidences for different categories. Inspired by the work (Hang and Zhang 2021), we here utilize label semantic features to perform feature relevance selection. Specifically, we multiply the activated $\mathbf{L}$ by the Sigmoid function element-wise with each instance's embedding feature $\hat{z}_i^{(v)}$ to obtain a new instance-label embedding matrix $\mathbf{B}_i^{(v)} \in \mathbb{R}^{c\times d_e}$ in terms of i-th instance:

$$\mathbf{B}_i^{(v)} = [\sigma_S(h_1) \odot \hat{z}_i^{(v)}; \sigma_S(h_2) \odot \hat{z}_i^{(v)}; ...; \sigma_S(h_c) \odot \hat{z}_i^{(v)}], \tag{8}$$

where σ_S is the Sigmoid activation function. Next, we perform category prediction on $\{\mathbf{B}_i^{(v)}\}_{i=1}^n$ via a linear classifier to obtain the predicted logits $\mathbf{P}^{(v)} \in \mathbb{R}^{n\times c}$ for v-th view.

We know that the ultimate goal of multi-view learning is to obtain consistent prediction results. To achieve that, we employ the late fusion approach to get the consistent prediction of multiple views. It should be noted that the embedding features we completed are not extracted from real instances, so the reliability of these filled features must be considered

in the fusion process of multiple predictions. Recalling the attention scores in the embedding imputation, missing instances are not equally expressed by available instances according to the various intensity of attention. For example, some instances are farther away from other un-missing instances, and the corresponding attention scores are generally lower in $\bar{\mathbf{A}}$, for which we can set a lower confidence for such imputation features. To be specific, we compute the maximum original attention (without scaling) of each instance with other instances as a confidence score:

$$\mathbf{Q}_{i,v} = \max(\{\tau \log \bar{\mathbf{A}}_{i,j} \mathbf{W}_{i,v}\}_{j=1}^{n}), \tag{9}$$

where $\mathbf{Q} \in \mathbb{R}^{n \times m}$ denotes the confidence matrix whose element $\mathbf{Q}_{i,v}$ stores the confidence score of i-th instance in v-th view. Eq. (9) can be understood as finding the maximum value of i-th row of the unshrunk joint attention matrix. Of course, un-missing instances do not need to be approximated by other instances, so their confidence should be set to the constant 1 in the final confidence matrix $\mathbf{Q}'$:

$$\mathbf{Q}' = (1 - \mathbf{W}) \odot \mathbf{Q} + \mathbf{W} \tag{10}$$

Now, with $\mathbf{Q}'$ and $\{\mathbf{P}^{(v)}\}_{v=1}^{m}$, we can compute the fused prediction $\bar{\mathbf{P}}$:

$$\bar{\mathbf{P}}_{i,:} = \sigma_S\Big(\frac{\sum_{v=1}^{m} \mathbf{P}_{i,:}^{(v)} \mathbf{Q}_{i,v}}{\sum_{v=1}^{m} \mathbf{Q}_{i,v}}\Big). \tag{11}$$

Finally, we employ masked binary cross-entropy function to compute loss:

$$\begin{aligned} \mathcal{L} = -\frac{1}{\sum_{i,j} \mathbf{G}_{i,j}} \sum_{i=1}^{n} \sum_{j=1}^{c} \Big(& \mathbf{Y}_{i,j} \log(\bar{\mathbf{P}}_{i,j}) \\ & + (1 - \mathbf{Y}_{i,j}) \log(1 - \bar{\mathbf{P}}_{i,j}) \Big) \mathbf{G}_{i,j}. \end{aligned} \tag{12}$$

In Eq. (12), we introduce the missing label index matrix $\mathbf{G}$ to mask the unknown tag in the process of loss function.

Overview of Our Method

Reviewing the four parts of our method in this section, i.e., label semantic representation learning, attention-induced missing view imputation, multi-view late-fusion and multi-label classification, we conclude as follows: (1) Our framework consists of 2 branches, an instance feature extraction branch and a label semantic extraction branch, which are fused in the embedding space to get the prediction of the sample on each view. (2) We adopt a multi-view late fusion method to fuse the prediction values of multiple views according to the confidence, and calculate the multi-label classification loss on the basis of available labels. (3) During the representation learning process, we approximate missing instances based on the available inter-instance attention. Finally, our framework includes only one loss function, and Algorithm 1 shows the training process of our model.

Algorithm 1: Training process of AIMNet

Input: Incomplete multi-view data $\{\mathbf{X}^{(v)}\}_{v=1}^{m}$, missing-view index matrix $\mathbf{W} \in \{0,1\}^{n \times m}$, partial multi-label $\mathbf{Y} \in \mathbb{R}^{n \times c}$, missing-label index matrix $\mathbf{G} \in \{0,1\}^{n \times c}$.
Output: Prediction $\bar{\mathbf{P}}$.

Randomly initialize model parameters and set hyperparameters (τ, learning rate, and training epochs E).
Compute label correlation matrix $\mathbf{C}$ by Eq. (1).
for $t = 0; t < E; t++$ **do**
  Extract label semantic feature $\mathbf{L}$ by Eqs. (2) and (3).
  Extract instance embedding features $\{\mathbf{Z}^{(v)}\}_{v=1}^{m}$ by m MLPs $\{\Psi_v\}_{v=1}^{m}$, respectively.
  Compute attention matrices $\{\mathbf{A}^{(v)}\}_{v=1}^{m}$ by Eq. (4).
  Compute joint attention matrix $\bar{\mathbf{A}}^{(v)}$ by Eq. (5).
  Fill missing instances in the embedding space by Eqs. (6) and (7).
  Compute instance-label embedding features $\{\mathbf{B}_1^{(v)}, \mathbf{B}_2^{(v)}, ..., \mathbf{B}_n^{(v)}\}_{v=1}^{m}$ by Eq. (8).
  Obtain view-specific predictions $\{\mathbf{P}^{(v)}\}_{v=1}^{m}$ by a linear classifier.
  Compute confidence matrix $\mathbf{Q}'$ by Eqs. (9) and (10).
  Obtain multi-view fusion prediction $\bar{\mathbf{P}}$ by Eq. (11) and compute classification loss $\mathcal{L}$ by Eq. (12).
  Update network parameters.
end for

Experiments

Experimental Settings

Datasets: In line with prior research (Tan et al. 2018; Li and Chen 2022; Liu et al. 2023a), we select five widely recognized multi-view multi-label datasets for our experiments, i.e., Corel5k, Pascal07, ESPGame, IAPRTC12, and MIR-FLICKR. Each dataset includes six distinct features, i.e., GIST, HSV, DenseHue, DenseSift, RGB, and LAB. Statistics of the five datasets refer to supplementary materials.

Incomplete multi-view partial multi-label data preprocessing: Following existing works (Tan et al. 2018; Li and Chen 2022; Liu et al. 2023a), we manually generate incomplete multi-view partial multi-label data on the bias of aforementioned five complete multi-view multi-label datasets, to simulate real world missing situation. Specifically, we randomly mask 50% of the instances from each view as the missing entities, while ensuring to keep one available view per sample. Furthermore, we randomly select 70% of all data as the training set for comprehensive evaluation of our proposed method. Finally, for each category in training set, we randomly eliminate 50% of the positive and negative labels to generate partial multi-label data.

Comparison methods: In our experiments, we select eight state-of-the-art methods for comparison with our AIMNet. In the related works section, we have introduced iMVWL, NAIM3L, DICNet, GLOCAL, and CDMM. In addition to these five methods, we take three additional methods, namely C2AE (Yeh et al. 2017), DM2L (Ma and Chen 2021), and LVSL (Zhao et al. 2022a). C2AE is a deep neural network model that integrates canonical correlation analysis and autoencoder architectures for effective and robust multi-label classification. It is important to highlight that only iMVWL, NAIM3L, and DICNet have the capability to handle both incomplete views and labels. Consequently, following existing works (Tan et al. 2018; Li and

Data	Metric	C2AE	GLOCAL	CDMM	DM2L	LVSL	iMVWL	NAIM3L	DICNet	AIMNet
Corel5k	AP	$0.227_{0.008}$	$0.285_{0.004}$	$0.354_{0.004}$	$0.262_{0.005}$	$0.342_{0.004}$	$0.283_{0.008}$	$0.309_{0.004}$	$0.381_{0.004}$	$0.400_{0.010}$
	1-HL	$0.980_{0.002}$	$0.987_{0.000}$	$0.987_{0.000}$	$0.987_{0.000}$	$0.987_{0.000}$	$0.978_{0.000}$	$0.987_{0.000}$	$0.988_{0.000}$	$0.988_{0.000}$
	1-RL	$0.804_{0.010}$	$0.840_{0.003}$	$0.884_{0.003}$	$0.843_{0.002}$	$0.881_{0.003}$	$0.865_{0.005}$	$0.878_{0.002}$	$0.882_{0.004}$	$0.902_{0.002}$
	AUC	$0.806_{0.010}$	$0.843_{0.003}$	$0.888_{0.003}$	$0.845_{0.002}$	$0.884_{0.003}$	$0.868_{0.005}$	$0.881_{0.002}$	$0.884_{0.004}$	$0.905_{0.003}$
	1-OE	$0.246_{0.016}$	$0.327_{0.010}$	$0.410_{0.007}$	$0.295_{0.014}$	$0.391_{0.009}$	$0.311_{0.015}$	$0.350_{0.009}$	$0.468_{0.007}$	$0.475_{0.018}$
	1-Cov	$0.596_{0.016}$	$0.648_{0.006}$	$0.723_{0.007}$	$0.647_{0.005}$	$0.718_{0.006}$	$0.702_{0.008}$	$0.725_{0.005}$	$0.727_{0.011}$	$0.771_{0.005}$
	Ave.R	8.83	6.33	2.83	6.83	3.83	6.83	4.33	2.17	**1.00**
Pascal07	AP	$0.485_{0.008}$	$0.496_{0.004}$	$0.508_{0.005}$	$0.471_{0.008}$	$0.504_{0.005}$	$0.437_{0.018}$	$0.488_{0.003}$	$0.505_{0.012}$	$0.548_{0.008}$
	1-HL	$0.908_{0.002}$	$0.927_{0.000}$	$0.931_{0.001}$	$0.928_{0.001}$	$0.930_{0.000}$	$0.882_{0.004}$	$0.928_{0.001}$	$0.929_{0.001}$	$0.931_{0.001}$
	1-RL	$0.745_{0.009}$	$0.767_{0.004}$	$0.812_{0.004}$	$0.761_{0.005}$	$0.806_{0.003}$	$0.736_{0.015}$	$0.783_{0.001}$	$0.783_{0.008}$	$0.831_{0.004}$
	AUC	$0.765_{0.010}$	$0.786_{0.003}$	$0.838_{0.003}$	$0.779_{0.004}$	$0.832_{0.002}$	$0.767_{0.015}$	$0.811_{0.001}$	$0.809_{0.006}$	$0.851_{0.004}$
	1-OE	$0.438_{0.008}$	$0.443_{0.005}$	$0.419_{0.008}$	$0.420_{0.011}$	$0.419_{0.008}$	$0.362_{0.023}$	$0.421_{0.006}$	$0.427_{0.015}$	$0.461_{0.013}$
	1-Cov	$0.680_{0.010}$	$0.703_{0.004}$	$0.759_{0.003}$	$0.692_{0.004}$	$0.751_{0.003}$	$0.677_{0.015}$	$0.727_{0.002}$	$0.731_{0.006}$	$0.783_{0.004}$
	Ave.R	7.17	5.33	2.83	6.67	3.83	8.83	4.83	4.00	**1.00**
ESPGame	AP	$0.202_{0.006}$	$0.221_{0.002}$	$0.289_{0.003}$	$0.212_{0.002}$	$0.285_{0.003}$	$0.244_{0.005}$	$0.246_{0.002}$	$0.297_{0.002}$	$0.305_{0.004}$
	1-HL	$0.971_{0.002}$	$0.982_{0.000}$	$0.983_{0.000}$	$0.982_{0.000}$	$0.983_{0.000}$	$0.972_{0.000}$	$0.983_{0.000}$	$0.983_{0.000}$	$0.983_{0.000}$
	1-RL	$0.772_{0.006}$	$0.780_{0.004}$	$0.832_{0.001}$	$0.781_{0.001}$	$0.829_{0.001}$	$0.808_{0.002}$	$0.818_{0.002}$	$0.832_{0.001}$	$0.846_{0.002}$
	AUC	$0.777_{0.006}$	$0.784_{0.004}$	$0.836_{0.001}$	$0.785_{0.001}$	$0.833_{0.002}$	$0.813_{0.002}$	$0.824_{0.002}$	$0.836_{0.001}$	$0.850_{0.002}$
	1-OE	$0.262_{0.018}$	$0.317_{0.005}$	$0.396_{0.005}$	$0.294_{0.006}$	$0.389_{0.004}$	$0.343_{0.013}$	$0.339_{0.003}$	$0.439_{0.007}$	$0.442_{0.011}$
	1-Cov	$0.497_{0.011}$	$0.496_{0.006}$	$0.574_{0.004}$	$0.488_{0.003}$	$0.567_{0.005}$	$0.548_{0.004}$	$0.571_{0.003}$	$0.593_{0.003}$	$0.624_{0.005}$
	Ave.R	8.67	7.33	2.33	7.50	3.67	6.17	4.33	1.83	**1.00**
IAPRTC12	AP	$0.224_{0.007}$	$0.256_{0.002}$	$0.305_{0.004}$	$0.234_{0.003}$	$0.304_{0.004}$	$0.237_{0.003}$	$0.261_{0.001}$	$0.323_{0.001}$	$0.329_{0.005}$
	1-HL	$0.965_{0.002}$	$0.980_{0.000}$	$0.981_{0.000}$	$0.980_{0.000}$	$0.981_{0.000}$	$0.969_{0.000}$	$0.980_{0.000}$	$0.981_{0.000}$	$0.981_{0.000}$
	1-RL	$0.806_{0.005}$	$0.825_{0.002}$	$0.862_{0.002}$	$0.823_{0.002}$	$0.861_{0.002}$	$0.833_{0.002}$	$0.848_{0.001}$	$0.873_{0.001}$	$0.883_{0.003}$
	AUC	$0.807_{0.005}$	$0.830_{0.001}$	$0.864_{0.002}$	$0.825_{0.001}$	$0.863_{0.001}$	$0.835_{0.001}$	$0.850_{0.001}$	$0.874_{0.000}$	$0.885_{0.003}$
	1-OE	$0.300_{0.031}$	$0.378_{0.007}$	$0.432_{0.008}$	$0.340_{0.006}$	$0.429_{0.009}$	$0.352_{0.008}$	$0.390_{0.005}$	$0.468_{0.002}$	$0.459_{0.08}$
	1-Cov	$0.523_{0.009}$	$0.534_{0.003}$	$0.597_{0.004}$	$0.529_{0.004}$	$0.597_{0.004}$	$0.564_{0.005}$	$0.592_{0.004}$	$0.649_{0.001}$	$0.673_{0.006}$
	Ave.R	9.00	6.33	2.67	7.50	3.33	6.67	5.00	1.67	**1.17**
MIRFLICKR	AP	$0.505_{0.008}$	$0.537_{0.002}$	$0.570_{0.002}$	$0.514_{0.006}$	$0.553_{0.002}$	$0.490_{0.012}$	$0.551_{0.002}$	$0.589_{0.005}$	$0.602_{0.004}$
	1-HL	$0.853_{0.004}$	$0.874_{0.001}$	$0.886_{0.001}$	$0.878_{0.001}$	$0.885_{0.001}$	$0.839_{0.002}$	$0.882_{0.001}$	$0.888_{0.002}$	$0.890_{0.001}$
	1-RL	$0.821_{0.003}$	$0.832_{0.001}$	$0.856_{0.001}$	$0.831_{0.003}$	$0.856_{0.001}$	$0.803_{0.008}$	$0.844_{0.001}$	$0.863_{0.004}$	$0.873_{0.002}$
	AUC	$0.810_{0.004}$	$0.828_{0.001}$	$0.846_{0.001}$	$0.828_{0.003}$	$0.844_{0.001}$	$0.787_{0.012}$	$0.837_{0.001}$	$0.849_{0.004}$	$0.861_{0.001}$
	1-OE	$0.505_{0.020}$	$0.552_{0.005}$	$0.631_{0.004}$	$0.510_{0.008}$	$0.607_{0.004}$	$0.511_{0.022}$	$0.585_{0.003}$	$0.637_{0.007}$	$0.651_{0.006}$
	1-Cov	$0.590_{0.005}$	$0.605_{0.003}$	$0.640_{0.001}$	$0.604_{0.005}$	$0.636_{0.001}$	$0.572_{0.013}$	$0.631_{0.002}$	$0.652_{0.007}$	$0.671_{0.004}$
	Ave.R	8.17	6.17	3.00	6.83	3.83	8.67	5.00	2.00	**1.00**

Table 1: Experimental results of nine methods on the five datasets with 50% missing-view rate and 50% missing-label rate (the bottom right digit is the standard deviation). The average ranking on the six metrics is shown at 'Ave.R'.

Chen 2022), several modifications are made on the other five methods before conducting the experiments. Specifically, for those approaches can not cope with missing views (CDMM and LVSL, etc), we populate each view's missing instances with mean values of its available instances. For the approaches like DM2L and C2AE do not have the ability to process missing views, we treat the unknown labels as negative tags in these methods. Finally, considering that C2AE, GLOCAL, and DM2L are incomplete single-view multi-label classification methods, we conduct experiments on each view and select the best results as the multi-view classification results in our experiments.

Evaluation metrics: To be consistent with existing classic works (Tan et al. 2018; Li and Chen 2022), we adopt four popular performance metrics, i.e., ranking loss (RL), average precision (AP), Hamming loss (HL), and area under the adaptation curve (AUC) to evaluate our work. In addition to them, we also introduce two common multi-label classification metrics, OneError (OE) and Coverage (Cov) in the experiments. Notably, we calculate 1-RL, 1-HL, 1-OE, and 1-Cov as the final measure so that higher values indicate superior performance in all six metrics.

Experimental Results and Analysis

In this section, we compare our method with other eight advanced algorithms on the five datasets mentioned above and the experimental results of the six evaluation metrics are shown in Table 1, in which the missing-view rate and missing-label rate are both arranged as 50%. Additionally, we also report the average ranking of each method across all metrics in 'Ave.R'. According to Table 1, we can have the following observations:

- Of all five datasets, our AIMNet achieves overwhelming lead on most metrics. On the four datasets of Corel5k, Pascal07, ESPGame, and MIRFLICKR, AIMNet ranks 1st in all metrics, and on IAPRTC12, its average ranking is also as high as 1.17, which fully verifies the effectiveness of our method on iMvPMLC task.
- From Table 1, we can observe that methods represented by AIMNet and DICNet, which consider the incomplete views and partial labels at the same time, exhibit better

Method	Corel5k						ESPGame					
	AP	1-HL	1-RL	AUC	1-OE	1-Cov	AP	1-HL	1-RL	AUC	1-OE	1-Cov
AIMNet w/o LM	0.380	0.987	0.895	0.898	0.446	0.755	0.293	0.983	0.842	0.847	0.414	0.616
AIMNet w/o imp	0.390	0.987	0.896	0.898	0.457	0.764	0.294	0.982	0.840	0.844	0.427	0.608
AIMNet	0.400	0.988	0.902	0.905	0.475	0.771	0.305	0.983	0.846	0.850	0.442	0.624

Table 2: Ablation results on two datasets with 50% missing views and 50% missing labels. 'w/o' means 'without'.

Phase \ Method	C2AE	GLOCAL	CDMM	DM2L	LVSL	iMVWL	NAIML	DICNet	AIMNet
Training	170.24	154.42	16.02	713.37	63.73	165.82	143.63	313.89	244.12
Inference	0.04	0.89	1.73	0.04	0.64	0.02	0.01	0.05	0.03

Table 3: Time cost of training and inference phases on the Corel5k dataset with 70% training samples. (Unit: s)

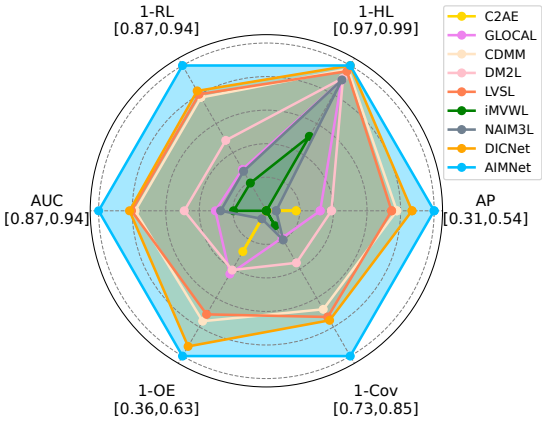

Figure 3: Experimental results of nine methods on the a full dataset Corel5k without any missing views or labels. The worst results are indicated at the center of radar map, while the best results are represented by the vertexes, considering six evaluation metrics.

performance than other methods in a single missing setting. This underscores the importance of addressing the challenges posed by missing views and labels.

- As late fusion methods, CDMM and AIMNet have both achieved higher rankings than other methods that directly obtain consistent results, which shows the effectiveness of multi-view decision level fusion to a certain extent.

Additionally, in order to further confirm the good adaptability of our model, we perform experiments on the complete dataset without any missing instances or labels. We present the results in Fig. 3 in radar charts (results on other datasets refer to supplementary material). Apparently, our AIMNet outperforms almost all other methods including those designed for ideal complete cases, demonstrating the generalization capability of our model.

Ablation Study

To evaluate the effectiveness of each component in our method, we perform ablation experiments on two datasets, i.e., Corel5k and ESPGame, where the proportions of missing views and missing labels are both 50%. Specifically, we employ two degradation methods, 'AIMNet w/o LM' and 'AIMNet w/o imp' for short. For 'AIMNet w/o LM', we simply remove the missing view index matrix $\mathbf{G}$ in the model, respectively. For 'AIMNet w/o imp', we remove the attention-induced view completion module and set corresponding confidence score as constant 0. The results of ablation experiments are listed in Table 2. It can be observed that the embedding feature imputation strategy make a significant contribution to the performance of model.

Time Cost Comparison

To study the training and inference efficiency of our AIMNet, we report the time cost of the training and test phases of the nine methods on the Corel5k dataset with 70% training samples in Table 3. It is well known that the training time of a model depends heavily on the setting of convergence conditions, so we measure the running times of all methods under their default convergence conditions. For special single-view methods, we record the total training time for all views and the inference time for a single view. From Table 3, we can observe that DNN based methods usually require more training time and less inference time.

Conclusion

In this paper, we propose a novel general framework (AIMNet) for the iMvPMLC task. The AIMNet is a two branch multi-view multi-label classification framework with a multi-view feature extraction module and a GAT based label semantic feature extraction module, modeling multi-label semantic information visually without any additional text pre-trained network. Additionally, instead of ignoring or skipping missing instances commonly used in existing works, our AIMNet complete the missing embedding features based on cross-view joint attention. During the multi-view late fusion phase, we develop a simple confidence based weighted fusion strategy to get the consistent classification results. Experimental results compared with eight state-of-art methods confirm the effectiveness and superiority of our proposed method.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grant No. 62372136, No. 62301621, and No. 62371157.

References

Chen, T.; Xu, M.; Hui, X.; Wu, H.; and Lin, L. 2019. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 522–531.

Fang, X.; Liu, D.; Zhou, P.; and Hu, Y. 2022. Multi-modal cross-domain alignment network for video moment retrieval. *IEEE Transactions on Multimedia*.

Fang, X.; Liu, D.; Zhou, P.; and Nan, G. 2023a. You Can Ground Earlier than See: An Effective and Efficient Pipeline for Temporal Sentence Grounding in Compressed Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2448–2460.

Fang, X.; Liu, D.; Zhou, P.; Xu, Z.; and Li, R. 2023b. Hierarchical local-global transformer for temporal sentence grounding. *IEEE Transactions on Multimedia*.

Fang, Z.; and Zhang, Z. 2012. Simultaneously Combining Multi-view Multi-label Learning with Maximum Margin Classification. In *2012 IEEE 12th International Conference on Data Mining*, 864–869.

Hang, J.-Y.; and Zhang, M.-L. 2021. Collaborative learning of label semantics and deep label-specific features for multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9860–9871.

Huang, J.; Li, G.; Huang, Q.; and Wu, X. 2015. Learning label specific features for multi-label classification. In *2015 IEEE International Conference on Data Mining*, 181–190. IEEE.

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*, 35(1): 589–602.

Li, X.; and Chen, S. 2022. A Concise Yet Effective Model for Non-Aligned Incomplete Multi-View and Missing Multi-Label Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 5918–5932.

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

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

Liu, C.; Wen, J.; Wu, Z.; Luo, X.; Huang, C.; and Xu, Y. 2023c. Information Recovery-Driven Deep Incomplete Multiview Clustering Network. *IEEE Transactions on Neural Networks and Learning Systems*, 1–11.

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

Liu, X.; Li, M.; Tang, C.; Xia, J.; Xiong, J.; Liu, L.; Kloft, M.; and Zhu, E. 2020. Efficient and effective regularized incomplete multi-view clustering. *IEEE transactions on pattern analysis and machine intelligence*, 43(8): 2634–2646.

Liu, X.; Zhu, X.; Li, M.; Wang, L.; Tang, C.; Yin, J.; Shen, D.; Wang, H.; and Gao, W. 2018. Late fusion incomplete multi-view clustering. *IEEE transactions on pattern analysis and machine intelligence*, 41(10): 2410–2423.

Liu, X.; Zhu, X.; Li, M.; Wang, L.; Zhu, E.; Liu, T.; Kloft, M.; Shen, D.; Yin, J.; and Gao, W. 2019. Multiple kernel k k-means with incomplete kernels. *IEEE transactions on pattern analysis and machine intelligence*, 42(5): 1191–1204.

Liu, Y.; Wang, J.; Huang, C.; Wang, Y.; and Xu, Y. 2023e. CIGAR: Cross-Modality Graph Reasoning for Domain Adaptive Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23776–23786.

Liu, Y.; Wang, J.; Xiao, L.; Liu, C.; Wu, Z.; and Xu, Y. 2023f. Foregroundness-Aware Task Disentanglement and Self-Paced Curriculum Learning for Domain Adaptive Object Detection. *IEEE Transactions on Neural Networks and Learning Systems*.

Luo, X.; Liu, C.; Wong, W.; Wen, J.; Jin, X.; and Xu, Y. 2023. MVCINN: multi-view diabetic retinopathy detection using a deep cross-interaction neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8993–9001.

Ma, Q.; Yuan, C.; Zhou, W.; and Hu, S. 2021. Label-specific dual graph neural network for multi-label text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3855–3864.

Ma, Z.; and Chen, S. 2021. Expand globally, shrink locally: Discriminant multi-label learning with missing labels. *Pattern Recognition*, 111: 107675.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2009. Classifier chains for multi-label classification. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II 20*, 254–269. Springer.

Tan, Q.; Yu, G.; Domeniconi, C.; and et al. 2018. Incomplete multi-view weak-label learning. In *Ijcai*, 2703–2709.

Trivedi, A.; Rai, P.; Daumé III, H.; and DuVall, S. L. 2010. Multiview clustering with incomplete views. In *NIPS workshop*, volume 224, 1–8. Citeseer.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. *International Conference on Learning Representations*. Accepted as poster.

Xu, C.; Guan, Z.; Zhao, W.; Wu, H.; Niu, Y.; and Ling, B. 2019. Adversarial incomplete multi-view clustering. In *IJCAI*, volume 7, 3933–3939.

Xu, J.; Li, C.; Ren, Y.; Peng, L.; Mo, Y.; Shi, X.; and Zhu, X. 2022a. Deep incomplete multi-view clustering via mining cluster complementarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8761–8769.

Xu, J.; Ren, Y.; Shi, X.; Shen, H. T.; and Zhu, X. 2023. UNTIE: Clustering analysis with disentanglement in multi-view information fusion. *Information Fusion*, 100: 101937.

Xu, J.; Ren, Y.; Tang, H.; Yang, Z.; Pan, L.; Yang, Y.; Pu, X.; Philip, S. Y.; and He, L. 2022b. Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*.

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.

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.

Zhao, D.; Gao, Q.; Lu, Y.; and Sun, D. 2022a. Non-Aligned Multi-View Multi-Label Classification Via Learning View-Specific Labels. *IEEE Transactions on Multimedia*.

Zhao, X.; Chen, Y.; Liu, S.; and Tang, B. 2022b. Shared-private memory networks for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*.

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