Partial Multi-View Multi-Label Classification via Semantic Invariance Learning and Prototype Modeling

Chengliang Liu [1] **Gehui Xu** [1] **Jie Wen** [1] **Yabo Liu** [1] **Chao Huang** [2] **Yong Xu** [1]

Abstract

The difficulty of partial multi-view multi-label learning lies in coupling the consensus of multi-view data with the task relevance of multi-label classification, under the condition where partial views and labels are unavailable. In this paper, we seek to compress cross-view representation to maximize the proportion of shared information to better predict semantic tags. To achieve this, we establish a model consistent with the information bottleneck theory for learning cross-view shared representation, minimizing non-shared information while maintaining feature validity to help increase the purity of task-relevant information. Furthermore, we model multi-label prototype instances in the latent space and learn label correlations in a data-driven manner. Our method outperforms existing state-of-the-art methods on multiple public datasets while exhibiting good compatibility with both partial and complete data. Finally, we experimentally reveal the importance of condensing shared information under the premise of information balancing, in the process of multi-view information encoding and compression.

1. Introduction

Throughout the years, we have been consistently pursuing the evaluation or decision-making of target objects based on multiple angles, levels, and diversities of information (Luo et al., 2023; Xiao et al., 2024; Luo et al., 2021; Liu et al., 2023d). In the field of data analysis and pattern recognition, multi-view learning is being used to enhance various machine learning tasks due to its powerful potential for representation learning (Wu & Goodman, 2018; Wang et al., 2023). Whether it is the classification or regression tasks based on multi-modal data with significant modal discrepancy, or the feature learning on multi-feature data originating from the same modality, they can all be unified under the theoretical framework of multi-view learning, which focuses on extracting and utilizing semantic invariance to facilitate downstream decision-making (Hwang et al., 2021; Zeng et al., 2023).

In recent years, traditional multi-view learning methods, especially multi-view representation learning methods, have been gradually replaced by deep multi-view learning networks, thanks to the powerful mapping capabilities of deep neural networks that surpass manual feature extraction (Wang et al., 2022; Liu et al., 2023e;f). For instance, some works employ deep autoencoders to extract view-specific features and uses contrastive learning to aggregate cross-view semantic representations in a high-level feature space (Xu et al., 2022; Liu et al., 2023a; 2024). Castrejon et al. attempted to learn a modality-independent shared representation for cross-modal scene images using convolutional neural networks (Castrejon et al., 2016). Wen et al. employed feature-weighted fusion to achieve consistent representation learning in the embedding space (Wen et al., 2023). In addition, Lee et al. proposed to learn task-relevant information while minimizing the mutual information between shared view representation and original data (Lee & Van der Schaar, 2021). It is evident that a core motivation of multi-view learning theory is to eliminate or reduce redundant information or uncertainty in single-view representation by utilizing and exploiting multi-view shared information. As research on multi-view learning advances, a more general and challenging scenario, in which partial views of some or all of the data are missing, has emerged as a new research hotspot, and we refer to tasks based on such settings as incomplete multi-view learning (Liu et al., 2022; 2023c).

Reflecting on the developmental trajectory of multi-view learning, we clearly observe that from unsupervised multi-view clustering tasks to supervised multi-view classification tasks, multi-view learning is evolving towards a close integration with downstream tasks. In this paper, our focus is on a specific and complex supervised task: multi-label

[1]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China [2]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen, China. Correspondence to: Jie Wen <jiewen_pr@126.com>, Yong Xu <yongxu@ymail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

classification. The combination of multi-view learning and multi-label classification greatly satisfies the demands of real-world application scenarios (Li & Chen, 2022).

Different from single-label learning, the important characteristic of multi-label classification is to model both the relationship between samples and labels, as well as the correlation among labels (Bai et al., 2022). Existing works in explicitly modeling label correlation can be categorized into the following strategies: (1) Calculating the conditional probability between category pairs based on the statistical information of the dataset as a priori knowledge of label correlation (You et al., 2020; Ma et al., 2021). (2) Dynamically learn label correlations in traditional convex optimization or neural network feedback training (Zhu et al., 2017). Another representative kind of method focuses on directly learning the mapping of each category in the latent space, trying to directly obtain classification results or enhance the discriminative ability of joint embeddings through the interaction between label and sample embeddings (Bai et al., 2022; Hang & Zhang, 2021). Similar to incomplete multi-view learning, label incompleteness is also incorporated into the problem setting to be more consistent with real situations.

Aiming at the above complex problem of partial multi-view multi-label classification (PMvMLC), our solution consists of two main parts: (1) Maximizing the proportion of shared information in the multi-view fusion representation. (2) Enhancing the task relevance of multi-label classification and multi-view shared information. Specifically, on the one hand, we assume that the shared information among views expresses all cross-view commonalities, including the highest level of semantic invariance (the description regarding semantic objectives should possess uniqueness across different views). Then, we take the variational autoencoder (VAE) as the framework of representation learning, and employ the information bottleneck theory to model the shared information of multi-view data, to approximate the ideal multi-view shared representation while minimizing the non-global shared private information of any view. On the other hand, we model multi-label prototype instances in the latent space to enhance the task relevance of the multi-label shared information. Specifically, we employ encoders to model the distribution of label prototypes, and then actively facilitate the integration of these label prototypes with samples' embedding representations, guided by prior supervisory information.

Overall, we name our **S**emantic **I**nvariance learning and **P**rototype modeling based method **SIP** and our contributions can be summarized as follows:

- We propose an information bottleneck based framework for PMvMLC that combines the learning of multi-view semantic invariance with the learning of multi-label prototype representations. Besides, our method can handle arbitrary view and label missing scenarios, demonstrating strong scalability.
- In contrast to existing works, we advocate that task-relevant information can be effectively compressed by extracting shared information across multiple views. Our proposed information bottleneck based framework effectively alleviates the challenges in extracting shared information from the partial multi-view data.
- Our method model the label prototypes in latent space via a data-driven approach, which effectively couples multi-view representation learning and multi-label classification tasks.
- Our method achieves leading performance on five complete or incomplete multi-view multi-label datasets, and the semantic invariance learning framework has good scalability for other multi-view learning tasks.

2. Preliminary

Given dataset containing n labeled samples $(\{\mathbf{x}^{(v)}\}_{v=1}^{m}, \mathbf{y})$ with m views, in which v-th view of any sample is $\mathbf{x}^{(v)} \in \mathbb{R}^{d_v}$ and corresponding label $\mathbf{y} \in \{0,1\}^c$ with c categories. Furthermore, we set $\mathcal{V}, |\mathcal{V}| \leq m$ as the observed view set and thus the multi-view data can be defined as $\{\mathbf{x}^{(v)}\}_{v\in\mathcal{V}}$ in the missing-view setting. Similarly, for the partial label setting, we let $\mathcal{U}, |\mathcal{U}| \leq c$ denotes the set of known tags. Our goal is to learn the cross-view representation $\mathbf{z}$ on $\{\mathbf{x}^{(v)}\}_{v\in\mathcal{V}}$ ($\{\mathbf{x}\}$ for short), and accurately predict the categories of $\{\mathbf{x}\}$ according to $\mathbf{z}$.

To achieve this, inspired by the work (Federici et al., 2020), we can give the following proposition:

Proposition 2.1. *If $\mathbf{z}$ holds all the information shared by all views, $\mathbf{z}$ can adequately predict $\mathbf{y}$.*

In other words, semantic information is included in the information shared by multiple views. Therefore, learning multi-view shared information has the opportunity to reduce the impact of individual view or partial views' private information on the prediction task while fully maintaining semantic information.

3. Method

3.1. Learning Shared Information via Information Bottleneck Principle

Proposition 2.1 explains that learning shared information is necessary to obtain task-relevant semantic information. On this basis, we further narrow the scope of $\mathbf{z}$ so that it only contains shared information across views, and we can show the following:

Corollary 3.1. *Let $\mathbf{z}$ contains and only contains the shard information of all available views, then conditional mutually information $I(\{\bar{\boldsymbol{x}}\};\boldsymbol{z}|\boldsymbol{x}^{(v)})=0$, where $\{\{\bar{\boldsymbol{x}}\},\boldsymbol{x}^{(v)}\}=\{\boldsymbol{x}\}$.*

The Corollary 3.1 gives the necessary condition to obtain the ideal $\mathbf{z}$. If we pursue the minimization of $I(\{\bar{\mathbf{x}}\};\mathbf{z}|\mathbf{x}^{(v)})$ alone, it will inevitably lead to the collapse of the information of $\mathbf{z}$. Therefore, considering that the information of cross-view representing $\mathbf{z}$ is derived from raw multi-view data $\{\mathbf{x}^{(v)}\}_{v\in\mathcal{V}}$, we build the goal of maximizing mutually information $I(\mathbf{x}^{(v)},\mathbf{z})$ and give the following constrainted optimization problem:

$$\begin{aligned}&\max \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} I(\mathbf{x}^{(v)};\mathbf{z})\\ &\textbf{s.t.},\ \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} I(\{\bar{\mathbf{x}}\};\mathbf{z}|\mathbf{x}^{(v)})=0\end{aligned} \tag{1}$$

Problem (1) can be converted to the following form using the Lagrange multiplier method:

$$\max \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}(I(\mathbf{x}^{(v)};\mathbf{z})-\beta I(\{\bar{\mathbf{x}}\};\mathbf{z}|\mathbf{x}^{(v)})) \tag{2}$$

where $\beta\geq 0$ is the Lagrange multiplier. Observing Eq. (2), it is formally consistent with the information bottleneck theory, β can also be regarded as the trade-off coefficient to balance the effectiveness and compactness of information. Furthermore, the former in Eq. (2) aims to maintain the amount of information learned from the original data in $\mathbf{z}$, and the latter aims to compress the information in $\mathbf{z}$ to exclude multi-view non-shared information. For the former term, we have the lower bound:

$$\begin{aligned}&I(\mathbf{x}^{(v)};\mathbf{z})\\ &=\int\int p(\mathbf{x}^{(v)},\mathbf{z})\log\frac{p(\mathbf{x}^{(v)}|\mathbf{z})}{p(\mathbf{x}^{(v)})}d\mathbf{x}^{(v)}d\mathbf{z}\\ &=\int p(\mathbf{x}^{(v)})\int p(\mathbf{z}|\mathbf{x}^{(v)})\log p(\mathbf{x}^{(v)}|\mathbf{z})d\mathbf{x}^{(v)}d\mathbf{z}+H(\mathbf{x}^{(v)})\\ &\geq\int p(\mathbf{x}^{(v)})\int p(\mathbf{z}|\mathbf{x}^{(v)})\log p(\mathbf{x}^{(v)}|\mathbf{z})d\mathbf{x}^{(v)}d\mathbf{z}\end{aligned} \tag{3}$$

where $p(\cdot)$ denotes the probability density function. Here, since Eq. (3) is intractable, we approximate $p(\mathbf{x}^{(v)}|\mathbf{z})$ using a stochastic decoder $q^v(\mathbf{x}^{(v)}|\mathbf{z})$ whose output can be denoted as $\hat{\mathbf{x}}^{(v)}$, and then we can get a new variational lower bound of $I(\mathbf{x}^{(v)};\mathbf{z})$ for maximization:

$$\begin{aligned}&I(\mathbf{x}^{(v)};\mathbf{z})\\ &\geq\int p(\mathbf{x}^{(v)})\int p(\mathbf{z}|\mathbf{x}^{(v)})\log p(\mathbf{x}^{(v)}|\mathbf{z})d\mathbf{x}^{(v)}d\mathbf{z}\\ &=\int p(\mathbf{x}^{(v)})\int p(\mathbf{z}|\mathbf{x}^{(v)})\log q^v(\mathbf{x}^{(v)}|\mathbf{z})d\mathbf{x}^{(v)}d\mathbf{z}+\\ &\quad\int p(\mathbf{x}^{(v)})\int p(\mathbf{z}|\mathbf{x}^{(v)})\log\frac{p(\mathbf{x}^{(v)}|\mathbf{z})}{q^v(\mathbf{x}^{(v)}|\mathbf{z})}d\mathbf{x}^{(v)}d\mathbf{z}\\ &\geq\mathbb{E}_{\mathbf{x}^{(v)}\sim p(\mathbf{x}^{(v)})}\Big[\int p(\mathbf{z}|\mathbf{x}^{(v)})\log q^v(\mathbf{x}^{(v)}|\mathbf{z})d\mathbf{z}\Big]\end{aligned} \tag{4}$$

For the latter term in Model (2), we have following optimization objective:

$$\begin{aligned}&I(\{\bar{\mathbf{x}}\};\mathbf{z}|\mathbf{x}^{(v)})\\ &=\int\int p(\{\mathbf{x}\},\mathbf{z})\log\frac{p(\{\mathbf{x}\},\mathbf{z})p(\mathbf{x}^{(v)})}{p(\{\mathbf{x}\})p(\mathbf{z},\mathbf{x}^{(v)})}d\{\mathbf{x}\}d\mathbf{z}\\ &=\int\int p(\{\mathbf{x}\},\mathbf{z})\log\frac{p(\mathbf{z}|\{\{\mathbf{x}\}\})}{p(\mathbf{z}|\mathbf{x}^{(v)})}d\{\mathbf{x}\}d\mathbf{z}\end{aligned} \tag{5}$$

Apparently, Eq. (5) is also computationally difficult. In our paper, we utilize two stochastic encoders to approximate the distribution, i.e., $r^v(\mathbf{z}|\mathbf{x}^{(v)})\approx p(\mathbf{z}|\mathbf{x}^{(v)})$. $r^v(\mathbf{z}|\mathbf{x}^{(v)}):=\mathcal{N}(f_\mu^v(\mathbf{x}^{(v)}),f_{\sigma^2}^v(\mathbf{x}^{(v)})\mathbf{I})$, where $f_\mu^v(\cdot)$ and $f_{\sigma^2}^v(\cdot)$ denote the encoders for the mean and variance in v-th view, respectively. $\mathbf{I}$ is the unit matrix. Then we can get following variational upper bound:

$$\begin{aligned}&I(\{\bar{\mathbf{x}}\};\mathbf{z}|\mathbf{x}^{(v)})\\ &=\int\int p(\{\mathbf{x}\},\mathbf{z})\log\frac{p(\mathbf{z}|\{\{\mathbf{x}\}\})r^v(\mathbf{z}|\mathbf{x}^{(v)})}{p(\mathbf{z}|\mathbf{x}^{(v)})r^v(\mathbf{z}|\mathbf{x}^{(v)})}d\{\mathbf{x}\}d\mathbf{z}\\ &=\int\int p(\{\mathbf{x}\},\mathbf{z})\log\frac{p(\mathbf{z}|\{\mathbf{x}\})}{r^v(\mathbf{z}|\mathbf{x}^{(v)})}d\{\mathbf{x}\}d\mathbf{z}+\\ &\quad\int\int p(\{\mathbf{x}\},\mathbf{z})\log\frac{r^v(\mathbf{z}|\mathbf{x}^{(v)})}{p(\mathbf{z}|\mathbf{x}^{(v)})}d\{\mathbf{x}\}d\mathbf{z}\\ &\leq\int p(\{\mathbf{x}\})D_{KL}(p(\mathbf{z}|\{\mathbf{x}\})\|r^v(\mathbf{z}|\mathbf{x}^{(v)}))d\{\mathbf{x}\}\\ &=\mathbb{E}_{\{\mathbf{x}\}\sim p(\{\mathbf{x}\})}[D_{KL}(p(\mathbf{z}|\{\mathbf{x}\})\|r^v(\mathbf{z}|\mathbf{x}^{(v)}))]\end{aligned} \tag{6}$$

where $D_{KL}(\cdot||\cdot)$ means the Kullback-Leibler divergence.

In combination with Eqs. (4) and (6), for Eq. (2), our optimization objective is naturally converted to minimize its upper bound:

$$\begin{aligned}\mathcal{L}_{IB}=\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\big[&-\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\{\mathbf{x}\})}\log q^v(\mathbf{x}^{(v)}|\mathbf{z})\\ &+\beta D_{KL}(p(\mathbf{z}|\{\mathbf{x}\})\|r^v(\mathbf{z}|\mathbf{x}^{(v)}))\big]\end{aligned} \tag{7}$$

With the optimization objective Eq. (7), an important question is how to obtain the joint posterior $p(\mathbf{z}|\{\mathbf{x}\})$ of multiple

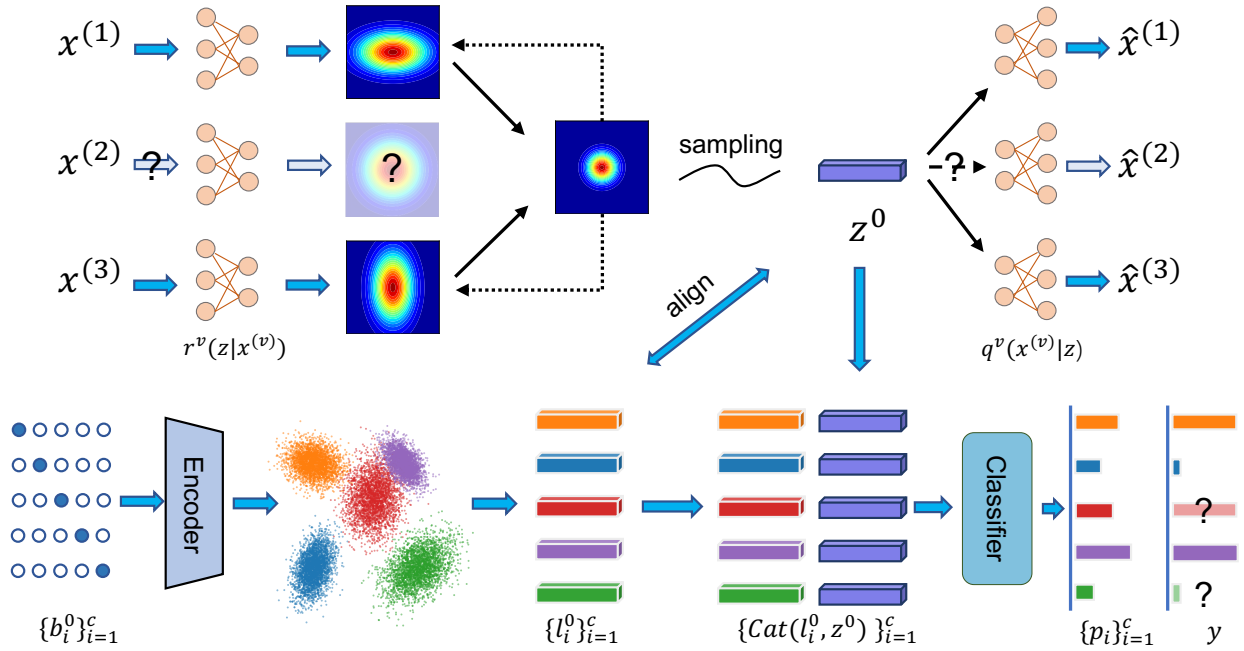


Figure 1: Our main framework of SIP. The upper part represents the partial multi-view shared information learning, and the lower part represents the data-driven label prototypes learning. In the upper part, we utilize the auto-encoders as the backbone, and the PoE category to fuse multiple views. In the lower part, label embeddings are extracted by a simple encoder and are aligned with the multi-view fusion feature.

views. Considering that we expect the shared information in the multi-view joint distribution to come from all views, i.e. each view provides the same shared information to $p(\mathbf{z}|\{\mathbf{x}\})$, we adopt the product-of-expert (PoE) (Hinton, 2002) with one-vote property to obtain the multi-view joint posterior. And then, following work (Wu & Goodman, 2018), we define the multi-view joint posterior as follows:

$$p(\mathbf{z}|\{\mathbf{x}\}) \propto p(\mathbf{z}) \prod_{v\in\mathcal{V}} p(\mathbf{z}|\mathbf{x}^{(v)}) := r(\mathbf{z}) \prod_{v\in\mathcal{V}} r^v(\mathbf{z}|\mathbf{x}^{(v)}) \quad (8)$$

where $r(\mathbf{z})$ is defined as a standard Gaussian distribution $r(\mathbf{z}) := \mathcal{N}(0, \mathbf{I})$ for a vanilla implementation. We expect the fusion view to show as much of the commonality of all views as possible. Moreover, the PoE fusion could preserve the probability distribution properties of representations, which is crucial for calculating our information theory based objective function.

3.2. Learning Label Prototype Representation

With the cross-view joint distribution, a simple way is to directly build a neural network as classifier to model the mapping of sample's latent space representation to the corresponding multi-label. However, as mentioned in the introduction, considering label correlation is an important difference between multi-label learning and single-label classification. Previous methods either use the pre-trained word-vector model to directly obtain label embeddings (Ma et al., 2021), or introduce prior label correlation to supervise the learning of label embeddings (You et al., 2020; Hang & Zhang, 2021). These methods lack flexibility and adaptability in modeling label-label relationship, especially in the case of missing labels. Motivated by the work (Bai et al., 2022), we tend to adopt a data-driven manner to learn the prototype representation of labels.

Similar to cluster centers, label prototypes are regarded as the projection of semantic targets in the embedding space, and well-trained label prototypes can help the model capture more discriminative information for multi-label classification. In this paper, to explicitly model the expression of the label's semantic concept, we attempt to deploy a stochastic encoder to fit the distribution of each category prototype, i.e., $\mathbf{l}_i \sim \mathcal{N}(\mu_i, \sigma_i^2\mathbf{I})$, where μ_i and σ_i^2 are the mean and variance output by the encoders $g_\mu(b_i)$ and $g_{\sigma^2}(b_i)$. $b_i \in \mathbb{R}^c$ is an available tensor initialized as an one-hot vector whose i-th bit is 1. Next, we sample s times from the distribution of $\mathbf{l}_i$ using the reparameterization trick to obtain the representation of label prototype in the latent space:

$$l_i^0 = \frac{1}{s}\sum_{d=1}^{s}(\mu_i + \sigma_i \odot \delta^d) \quad (9)$$

where δ^d means d-th sampling from the standard Gaussian distribution and $\odot$ denotes Hadamard product.

After initially modeling the distribution of each label prototype in the latent space, we further consider the correlation among labels. Generally, the occurrence of a particular label may indicate the co-occurrence of other labels that are associated with it and we define this phenomenon as label correlation. Introducing label correlation helps to avoid treating multi-label classification as simply multiple binary classification problems. However, manually designing and calculating label correlations based on statistical methods lacks flexibility, and thus we expect samples to play a dominant role in modeling label prototypes. Intuitively, a label prototype should represent the geometric center of samples belonging to that category. To do this, we enforce the label prototypes to move closer to their corresponding samples by minimizing the distance between prototypes and samples in the Euclidean space. Specifically, we sample the cross-view representation z^0 like Eq. (9) and its corresponding label prototypes $\{l_j^0|j \in \mathcal{C}\}$, where $\mathcal{C}$ means the set of positive known labels belonging to z^0, and then minimize their distance in the Euclidean space:

$$\mathcal{L}_{PA} = \| \frac{1}{|\mathcal{C}|} \sum_{j \in \mathcal{C}} l_j^0 - z^0 \|_2^2 \tag{10}$$

where $\mathcal{L}_{PA}$ is our prototype aggregation loss. According to Eq. (10), the label prototypes that appear at the same time will be close to the cross-view representation of the sample, which indirectly promotes the correlation learning between label prototypes.

3.3. Multi-Label Classification and Objective Function

The purpose of learning discriminative label prototypes is to obtain better classification results. Existing works predict the probability that a sample belongs to each category by calculating the inner product of label and sample embeddings (Bai et al., 2022). However, we argue that manually specifying the sample-label distance measure relies too much on experience. In this paper, we propose to measure the similarity between the sample and label prototypes via a simple neural network:

$$p_i = \omega(f_c(z^0 \circledcirc l_i^0)) \tag{11}$$

where $\circledcirc$ denotes concatenation operation, f_c is a sample fully connected layer, and $\omega(\cdot)$ is the $Sigmoid$ activation function. Finally, we get the prediction probability p_i that the sample is labeled as i-th category. Of course, our objective function also includes a multi-label cross-entropy loss, which endows the entire multi-label learning model with the maximum task relevance. The cross-entropy loss $\mathcal{L}_{CE}$ for each sample-prediction pair is expressed as:

$$\mathcal{L}_{CE} = \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} [y_i \log p_i + (1 - y_i) \log(1 - p_i)] \tag{12}$$

where y_i means the positive or negative mask of i-th category in $\mathbf{y}$. Considering that some labels are unknown, we ignore the items corresponding missing tags when calculating the cross-entropy loss. This simple and effective way can reduce the impact of label uncertainty on task correlation learning as much as possible.

At this point, we simply add the objective functions of our various parts to get the following overall optimization objective:

$$\mathcal{L} = \mathcal{L}_{IB} + \alpha \mathcal{L}_{PA} + \mathcal{L}_{CE} \tag{13}$$

where α is a trade-off coefficient and our loss function is defined in single sample case.

Review the above three parts, i.e., shared information learning, label prototype learning, and multi-label classification, in which the shared information learning is self-supervised and other parts are supervised learning. A natural idea is to conduct the self-supervised learning as a pre-training procedure for supervised learning (Khosla et al., 2020). However, in our scenario, the shared information obtained by self-supervised learning closely acts on task-relevant label prototype learning, so executing these two types of tasks simultaneously in a joint framework will not cause significant performance degradation, while maintaining better reproducibility.

Algorithm 1 shows the training process of the proposed mode. Note that during training, we adjust the network parameters by minimizing the objective function, including the parameters $\{b_i\}_{i=1}^c$ used to generate the label prototypes, while in the test phase, our label prototypes do not update as the input changes.

4. Experiments

4.1. Experimental Settings

Datasets: In line with previous works (Liu et al., 2023a; Tan et al., 2018; Li & Chen, 2022; Liu et al., 2023b), we adopt five widely recognized multi-view multi-label databases in our experiments, i.e., Corel5k (Duygulu et al., 2002), Pascal07 (Everingham et al., 2009), ESPGame (Von Ahn & Dabbish, 2004), IAPRTC12 (Henning et al., 2006), and MIRFLICKR (Huiskes & Lew, 2008). There are six distinct features, i.e., GIST, HSV, DenseHue, DenseSift, RGB, and LAB, in the five databases. More information of the five datasets refers to the appendix. **Incomplete multi-view partial multi-label data preprocessing:** Following existing works (Tan et al., 2018; Li & Chen, 2022; Liu et al., 2023a), to simulate missing data setting in real scenario, we need to generate partial multi-view multi-label data based on the five mentioned complete multi-view multi-label datasets. The process involves randomly disabling 50% of the instances from each view, ensuring that at least one view remained available, and randomly eliminating 50%

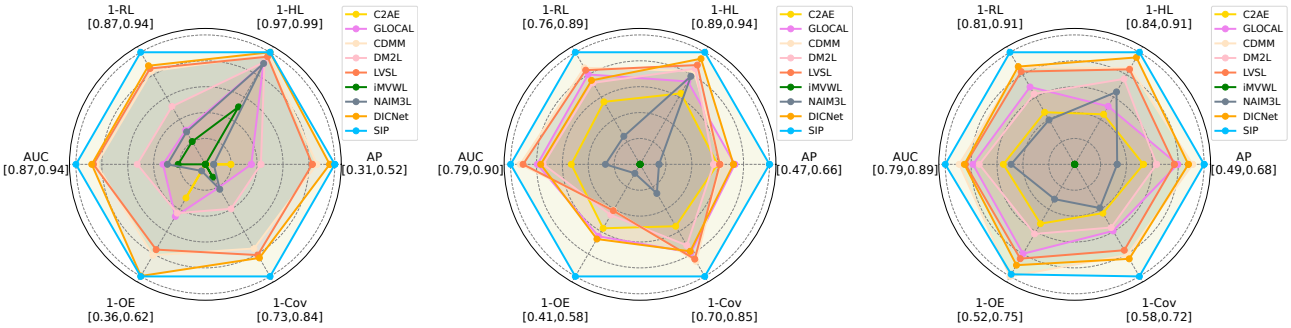
Table 1: Experimental results of nine methods on the five datasets with 50% missing-view rate and 50% missing-label rate (the bottom right digit is the standard deviation). The average ranking on the six metrics is shown at 'Ave.R'.

Data	Metric	C2AE	GLOCAL	CDMM	DM2L	LVSL	iMVWL	NAIM3L	DICNet	SIP
Corel5k	AP	$0.227_{0.008}$	$0.285_{0.004}$	$0.354_{0.004}$	$0.262_{0.005}$	$0.342_{0.004}$	$0.283_{0.008}$	$0.309_{0.004}$	$0.381_{0.004}$	$0.418_{0.009}$
	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	$0.804_{0.010}$	$0.840_{0.003}$	$0.884_{0.003}$	$0.843_{0.002}$	$0.881_{0.003}$	$0.865_{0.005}$	$0.878_{0.002}$	$0.882_{0.004}$	$0.911_{0.003}$
	AUC	$0.806_{0.010}$	$0.843_{0.003}$	$0.888_{0.003}$	$0.845_{0.002}$	$0.884_{0.003}$	$0.868_{0.005}$	$0.881_{0.002}$	$0.884_{0.004}$	$0.913_{0.003}$
	1-OE	$0.246_{0.016}$	$0.327_{0.010}$	$0.410_{0.007}$	$0.295_{0.014}$	$0.391_{0.009}$	$0.311_{0.015}$	$0.350_{0.009}$	$0.468_{0.007}$	$0.489_{0.016}$
	1-Cov	$0.596_{0.016}$	$0.648_{0.006}$	$0.723_{0.007}$	$0.647_{0.005}$	$0.718_{0.006}$	$0.702_{0.008}$	$0.725_{0.005}$	$0.727_{0.011}$	$0.787_{0.009}$
	Ave.R	8.83	6.33	2.83	6.83	3.83	6.83	4.33	2.17	**1.00**
Pascal07	AP	$0.485_{0.008}$	$0.496_{0.004}$	$0.508_{0.005}$	$0.471_{0.008}$	$0.504_{0.005}$	$0.437_{0.018}$	$0.488_{0.003}$	$0.505_{0.012}$	$0.555_{0.010}$
	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	$0.745_{0.009}$	$0.767_{0.004}$	$0.812_{0.004}$	$0.761_{0.005}$	$0.806_{0.003}$	$0.736_{0.015}$	$0.783_{0.001}$	$0.783_{0.008}$	$0.830_{0.004}$
	AUC	$0.765_{0.010}$	$0.786_{0.003}$	$0.838_{0.003}$	$0.779_{0.004}$	$0.832_{0.002}$	$0.767_{0.015}$	$0.811_{0.001}$	$0.809_{0.006}$	$0.850_{0.005}$
	1-OE	$0.438_{0.008}$	$0.443_{0.005}$	$0.419_{0.008}$	$0.420_{0.011}$	$0.419_{0.008}$	$0.362_{0.023}$	$0.421_{0.006}$	$0.427_{0.015}$	$0.464_{0.018}$
	1-Cov	$0.680_{0.010}$	$0.703_{0.004}$	$0.759_{0.003}$	$0.692_{0.004}$	$0.751_{0.003}$	$0.677_{0.015}$	$0.727_{0.002}$	$0.731_{0.006}$	$0.783_{0.006}$
	Ave.R	7.17	5.33	2.83	6.67	3.83	8.83	4.83	4.00	**1.00**
ESPGame	AP	$0.202_{0.006}$	$0.221_{0.002}$	$0.289_{0.003}$	$0.212_{0.002}$	$0.285_{0.003}$	$0.244_{0.005}$	$0.246_{0.002}$	$0.297_{0.002}$	$0.311_{0.004}$
	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	$0.772_{0.006}$	$0.780_{0.004}$	$0.832_{0.001}$	$0.781_{0.001}$	$0.829_{0.001}$	$0.808_{0.002}$	$0.818_{0.002}$	$0.832_{0.001}$	$0.849_{0.002}$
	AUC	$0.777_{0.006}$	$0.784_{0.004}$	$0.836_{0.001}$	$0.785_{0.001}$	$0.833_{0.002}$	$0.813_{0.002}$	$0.824_{0.002}$	$0.836_{0.001}$	$0.853_{0.002}$
	1-OE	$0.262_{0.018}$	$0.317_{0.005}$	$0.396_{0.005}$	$0.294_{0.006}$	$0.389_{0.004}$	$0.343_{0.013}$	$0.339_{0.003}$	$0.439_{0.007}$	$0.455_{0.007}$
	1-Cov	$0.497_{0.011}$	$0.496_{0.006}$	$0.574_{0.004}$	$0.488_{0.003}$	$0.567_{0.005}$	$0.548_{0.004}$	$0.571_{0.003}$	$0.593_{0.003}$	$0.628_{0.005}$
	Ave.R	8.67	7.33	2.33	7.50	3.67	6.17	4.33	1.83	**1.00**
IAPRTC12	AP	$0.224_{0.007}$	$0.256_{0.002}$	$0.305_{0.004}$	$0.234_{0.003}$	$0.304_{0.004}$	$0.237_{0.003}$	$0.261_{0.001}$	$0.323_{0.001}$	$0.331_{0.006}$
	1-HL	$0.965_{0.002}$	$0.980_{0.000}$	$0.981_{0.000}$	$0.980_{0.000}$	$0.981_{0.000}$	$0.969_{0.000}$	$0.980_{0.000}$	$0.981_{0.000}$	$0.980_{0.000}$
	1-RL	$0.806_{0.005}$	$0.825_{0.002}$	$0.862_{0.002}$	$0.823_{0.002}$	$0.861_{0.002}$	$0.833_{0.002}$	$0.848_{0.001}$	$0.873_{0.001}$	$0.885_{0.003}$
	AUC	$0.807_{0.005}$	$0.830_{0.001}$	$0.864_{0.002}$	$0.825_{0.001}$	$0.863_{0.001}$	$0.835_{0.001}$	$0.850_{0.001}$	$0.874_{0.000}$	$0.886_{0.002}$
	1-OE	$0.300_{0.031}$	$0.378_{0.007}$	$0.432_{0.008}$	$0.340_{0.006}$	$0.429_{0.009}$	$0.352_{0.008}$	$0.390_{0.005}$	$0.468_{0.002}$	$0.463_{0.009}$
	1-Cov	$0.523_{0.009}$	$0.534_{0.003}$	$0.597_{0.004}$	$0.529_{0.004}$	$0.597_{0.004}$	$0.564_{0.005}$	$0.592_{0.004}$	$0.649_{0.001}$	$0.675_{0.007}$
	Ave.R	9.00	6.33	2.67	7.50	3.33	6.67	5.00	1.83	**1.00**
MIRFLICKR	AP	$0.505_{0.008}$	$0.537_{0.002}$	$0.570_{0.002}$	$0.514_{0.006}$	$0.553_{0.002}$	$0.490_{0.012}$	$0.551_{0.002}$	$0.589_{0.005}$	$0.614_{0.004}$
	1-HL	$0.853_{0.004}$	$0.874_{0.001}$	$0.886_{0.001}$	$0.878_{0.001}$	$0.885_{0.001}$	$0.839_{0.002}$	$0.882_{0.001}$	$0.888_{0.002}$	$0.891_{0.001}$
	1-RL	$0.821_{0.003}$	$0.832_{0.001}$	$0.856_{0.001}$	$0.831_{0.003}$	$0.856_{0.001}$	$0.803_{0.008}$	$0.844_{0.001}$	$0.863_{0.004}$	$0.877_{0.002}$
	AUC	$0.810_{0.004}$	$0.828_{0.001}$	$0.846_{0.001}$	$0.828_{0.003}$	$0.844_{0.001}$	$0.787_{0.012}$	$0.837_{0.001}$	$0.849_{0.004}$	$0.860_{0.003}$
	1-OE	$0.505_{0.020}$	$0.552_{0.005}$	$0.631_{0.004}$	$0.510_{0.008}$	$0.607_{0.004}$	$0.511_{0.022}$	$0.585_{0.003}$	$0.637_{0.007}$	$0.662_{0.008}$
	1-Cov	$0.590_{0.005}$	$0.605_{0.003}$	$0.640_{0.001}$	$0.604_{0.005}$	$0.636_{0.001}$	$0.572_{0.013}$	$0.631_{0.002}$	$0.652_{0.007}$	$0.678_{0.003}$
	Ave.R	8.17	6.17	3.00	6.83	3.83	8.67	5.00	2.00	**1.00**

of the positive and negative tags of each sample to generate the partial multi-label data. The missing views and labels are filled with '0' value to keep the dimensions unchanged. Finally, we randomly divide 70% of the entire dataset as the training set to comprehensively evaluate our proposed method.

Comparison methods: In our experiments, we select eight top approaches for comparison with our SIP, namely C2AE (Yeh et al., 2017), GLOCAL (Zhu et al., 2017), CDMM (Zhao et al., 2021), DM2L (Ma & Chen, 2021), LVSL (Zhao et al., 2022), iMVWL (Tan et al., 2018), NAIM3L (Li & Chen, 2022), and DICNet (Liu et al., 2023a). It must be emphasized that due to the complexity of our problem setup, not all of the methods introduced for comparison are perfect for missing multi-view and partial multi-label settings, such as C2AE, GLOCAL, CDMM, DM2L, and LVSL. To be specific, C2AE, GLOCAL, and DM2L can only handle the single-view partial multi-label case. CDMM and LVSL are complete multi-view multi-label classification models without any missing data processing ability. For above five methods, we have to make some changes to adept them to our setup: (1) We conduct experiments and record the best results of each view in single-view methods. (2) We fill the unavailable instances with mean values of corresponding view's available instances for those methods designed for full multi-view data. (3) We simply regard the unknown tag as the the negative tag for those methods unable to partial multi-label data. See the Appendix for more information on the comparison methods.

Evaluation metrics: Similar to previous works (Tan et al., 2018; Li & Chen, 2022), in our experiments, we adopt six popular performance metrics, namely ranking loss (RL), average precision (AP), Hamming loss (HL), area under the adaptation curve (AUC), OneError (OE), and Coverage (Cov) to evaluate our SIP. Note that we record 1-RL, 1-HL, 1-OE, and 1-Cov as the final results so that higher values mean superior performance in all six metrics.



(a) Corel5k (b) Pascal07 (c) MIRFLICKR

Figure 2: Experimental results of nine methods on three full datasets without any missing views or labels. The worst results are indicated at the center of radar chart, while the best results are represented by the vertexes, considering six evaluation metrics.

Algorithm 1 Training process of the proposed model

Input: Incomplete multi-view data $(\{\mathbf{x}^{(v)}\}_{v=1}^{m}, \mathbf{y})$, observed view set $\mathcal{V}$, known label set $\mathcal{U}$, training epochs k , and trade-off coefficients α, β.
Output: The parameters of model.
Initialize the parameters $\{b_i\}_{i=1}^{c}$ as $\mathbf{I} \in \{0,1\}^{c\times c}$.
for $t = 1, \ldots, k$ **do**
  Compute the conditional distribution of cross-view representation $\{r^v(\mathbf{z}|\mathbf{x}^{(v)})|v \in \mathcal{V}\}$ on each available view by encoders $f_{\mu}^{v}(\mathbf{x}^{(v)})$ and $f_{\sigma^2}^{v}(\mathbf{x}^{(v)})$.
  Obtain the distribution of cross-view representation $\mathbf{z}$ by Eq. (8).
  Sample z^0 from distribution of $\mathbf{z}$ like Eq. (9).
  Compute conditional distribution $\{q^v(\mathbf{x}^{(v)}|\mathbf{z})|v \in \mathcal{V}\}$ by z^0 and decoders.
  Compute c conditional distribution of prototype representation $\{\mathbf{l}_i \sim \mathcal{N}(g_{\sigma^2}^{v}(b_i), g_{\sigma^2}^{v}(b_i)\mathbf{I})|i = 1, ..., c\}$.
  Sample each l_i^0 from corresponding distribution of $\mathbf{l}_i$ by Eq. (9).
  Compute total loss: $\mathcal{L} = \mathcal{L}_{IB} + \alpha\mathcal{L}_{PA} + \mathcal{L}_{CE}$.
  Update the parameters of the model with $\mathcal{L}$;
end for

4.2. Experimental Results and Analysis

We compare our SIP with other nine top methods on the five datasets and show experimental results of the six evaluation metrics in Table 1, where the incompleteness rate of views and labels are both set as 50%. For a more intuitive comparison, we also calculate the average ranking of each method on six metrics ('Ave.R'). From Table 1, we can give following observations:

- Compared to the other nine competitors, our SIP achieves the best performance on all metrics. SIP ranked first in all datasets, fully verifying its effectiveness on the PMvMLC task.
- We can observe that DICNet and SIP that can handle incomplete views and partial labels perform better than others applied only work with missing label or ideal scenario. This provides insights for the design of multi-view multi-label classification models in the future.

In the radar chart Fig. 2, to further confirm that our model has good adaptability to complete multi-view data, we provide the results of nine methods on three datasets with full views and labels (refer to Appendix for more results on other two databases). Clearly, our SIP still achieves competitive performance than other methods, including those designed for the ideal complete case, demonstrating the generalization ability of our model.

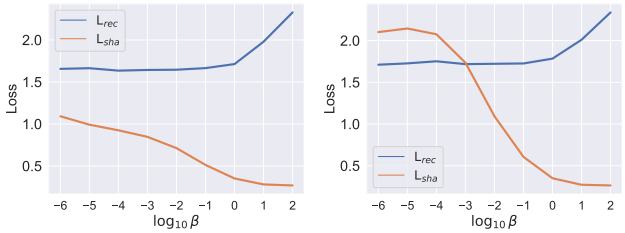
4.3. Analysis of the Balance in Information Bottleneck Modeling

Although the overall effectiveness of our SIP is confirmed in the comparison experiments, in order to further investigate the mechanism of information bottleneck principle on our task, we adjust the hype-parameter β in Eq. (7) to study the effect of each component on cross-view shard information. For convenience, we name the first part of Eq. (7) as $\mathcal{L}_{rec}$ and the second part as $\mathcal{L}_{sha}$. Fig. 3 depicts the curves of AP value and two losses ($\mathcal{L}_{rec}$ and $\mathcal{L}_{sha}$) as they vary with β over the same training epochs on the Corel5k and Pascal07 datasets with half available views and labels.

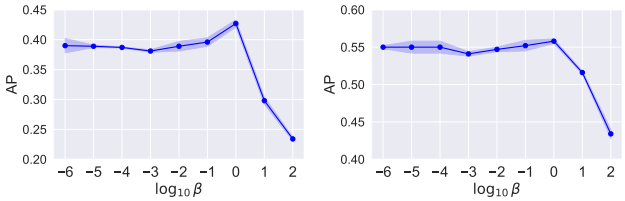
A clear trend can be seen from the Fig. 3, that is, as the β increases, the $\mathcal{L}_{rec}$ gradually increases and the $\mathcal{L}_{sha}$ gradually decreases. Reviewing the two components of our $\mathcal{L}_{IB}$, the first term is responsible for ensuring that the cross-view representation $\mathbf{z}$ is relevant to the raw data, and the second term is dedicated to constraining the learned $\mathbf{z}$ to contain only

Table 2: Ablation results on two datasets with 50% missing views and 50% missing labels. 'w/o' means 'without'. $\mathcal{L}_{sha}$ and $\mathcal{L}_{rec}$ denote two terms of $\mathcal{L}_{IB}$.

Method	Corel5k						Pascal07					
	AP	1-HL	1-RL	AUC	1-OE	1-Cov	AP	1-HL	1-RL	AUC	1-OE	1-Cov
SIP w/o $\mathcal{L}_{sha}$ $\mathcal{L}_{rec}$ $\mathcal{L}_{PA}$	0.345	0.987	0.874	0.877	0.423	0.729	0.535	0.931	0.813	0.836	0.454	0.763
SIP w/o $\mathcal{L}_{sha}$ $\mathcal{L}_{PA}$	0.383	0.987	0.898	0.901	0.454	0.761	0.547	0.929	0.825	0.847	0.464	0.777
SIP w/o $\mathcal{L}_{rec}$ $\mathcal{L}_{PA}$	0.387	0.988	0.873	0.876	0.469	0.726	0.511	0.931	0.786	0.813	0.440	0.736
SIP w/o $\mathcal{L}_{PA}$	0.415	0.987	0.912	0.912	0.487	0.790	0.549	0.931	0.826	0.847	0.461	0.778
SIP w/o $\mathcal{L}_{rec}$	0.257	0.987	0.853	0.855	0.290	0.678	0.516	0.931	0.792	0.817	0.443	0.742
SIP w/o $\mathcal{L}_{sha}$	0.383	0.987	0.898	0.900	0.467	0.762	0.550	0.930	0.827	0.848	0.463	0.780
SIP	0.418	0.988	0.911	0.913	0.489	0.787	0.555	0.931	0.830	0.850	0.464	0.783

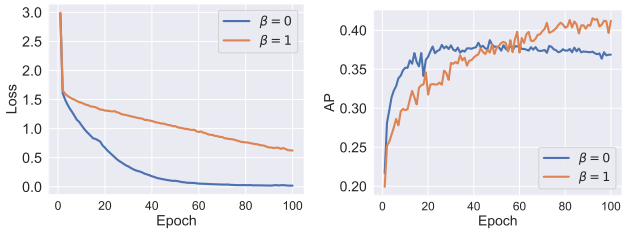


(a) Loss curve on the Corel5k (b) Loss curve on the Pascal07



(c) AP values on the Corel5k (d) AP values on the Pascal07

Figure 3: Impact of different information-balance parameter β on loss $\mathcal{L}_{rec}$, $\mathcal{L}_{sha}$ (3a,3b), and AP (3c, 3d). The blue area shows the standard deviation.



(a) Cross entropy on training set (b) AP values of test set

Figure 4: Impact of mining shared information on the training of SIP, shown by the cross-entropy (4a) and AP values (4b) on the Corel5k with 50% known views and labels.

shared information of available views as much as possible. When the β is small, the cross-view representation learned by the model contains a large amount of view-unshared information, which is quite beneficial for reconstructing the raw data. When we gradually increase β, it can effectively increase the proportion of shared information in $\mathbf{z}$, resulting a reduced $\mathcal{L}_{sha}$. When β is too large, the model focuses too much on reducing the non-shared information in $\mathbf{z}$, making it difficult for $\mathbf{z}$ to retain enough information for reconstruction, which corresponds to the sudden rise of $\mathcal{L}_{rec}$ in Fig. 3a and Fig. 3b after β is greater than $1e0$. Combined with Fig. 3c and Fig. 3d, when $\beta = 1e0$, the model achieves a balance between minimizing the reconstruction loss and maximizing the proportion of shared information, resulting in the optimal performance.

4.4. Analysis of Mining Shared Information

In Section 4.3, we discuss the balance between compressing and maintaining information in the information bottleneck method. Here, we try to remove shared information learning and observe what happens during the training process of the model. Specifically, we set β to 0 and 1 respectively during training, and then record the cross-entropy loss on the training set and AP value on the test set after each training epoch to analyze the impact of β on the model classification performance. Fig. 4 shows the experimental results on Corel5k dataset with 50% missing views and labels.

From the figure, learning compact shared information makes the model's classification loss during the training phase converge slower compared to uncompressed. However, although non-shared redundant information accelerates the fitting of training set, it shows obvious overfitting on the test set. This phenomenon supports our idea that learning shared information can improve the model's ability to extract high-level semantic information.

4.5. Ablation Study

To evaluate the effectiveness of each component of our SIP, we conduct extensive ablation experiments on two databases, namely Corel5k and Pascal07, in which the available rates of views and labels are both set as 50%. Our objective function consists of two parts, i.e., $\mathcal{L}_{IB}$ and $\mathcal{L}_{PA}$. For $\mathcal{L}_{IB}$, we split its two items into two parts for ablation experiments, namely $\mathcal{L}_{IB} = \mathcal{L}_{sha} + \beta\mathcal{L}_{rec}$. The results of ablation experiments are listed in Table 2. It can be observed that the two components of $\mathcal{L}_{IB}$ make a significant contribution to the

performance of model. For different databases, $\mathcal{L}_{sha}$ and $\mathcal{L}_{rec}$ show different influence on performance due to the varying difficulty of exploiting shared information, while maintaining the balance of effective information. Additionally, we can observe an interesting phenomenon that when $\mathcal{L}_{rec}$ is removed, adding $\mathcal{L}_{sha}$ does not improve the performance but rather leads to a significant performance decrease. This is because the compression of shared information without ensuring the effective information will cause $\mathbf{z}$ to lack enough information for prediction.

5. Conclusion

In this paper, we assume that the shared information among views encompasses all cross-view commonalities, including the highest level of semantic invariance. The descriptions related to semantic objectives should exhibit uniqueness across different views. Based on this assumption, we propose a general semantic invariance learning approach called SIP for PMvMLC task, which demonstrates good compatibility with missing views and incomplete labels. Our method divides the problem of multi-view multi-label learning into two parts: cross-view representation learning and multi-label prototype learning. To begin with, holding the idea that shared information can provide sufficient conditions for predicting semantic tasks, we utilize the information bottleneck theory to learn effective information while minimizing non-shared information in the cross-view representation, thereby maximizing the shared information across all views. Additionally, we attempt to drive the learning of label prototypes using the latent representation of the data, explicitly modeling label correlations. Finally, we conduct extensive experiments to validate the effectiveness of our method, while also discussing in-depth that the key to the success of our method lies in achieving a balance between extracting multi-view information and compressing shared information.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62372136 and No. 62301621), in part by Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515030213).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Bai, J., Kong, S., and Gomes, C. P. Gaussian mixture variational autoencoder with contrastive learning for multi-label classification. In *International Conference on Machine Learning*, pp. 1383–1398. PMLR, 2022.

Castrejon, L., Aytar, Y., Vondrick, C., Pirsiavash, H., and Torralba, A. Learning aligned cross-modal representations from weakly aligned data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2940–2949, 2016.

Duygulu, P., Barnard, K., de Freitas, J. F., and Forsyth, D. A. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European Conference on Computer Vision*, pp. 97–112, 2002.

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

Federici, M., Dutta, A., Forré, P., Kushman, N., and Akata, Z. Learning robust representations via multi-view information bottleneck. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1xwcyHFDr.

Hang, J.-Y. and Zhang, M.-L. 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, 2021.

Henning, M., Thomas, D., et al. The iapr benchmark: Anew evaluation resource for visual information systems. In *International Conference on Language Resources and Evaluation*, pp. 1–11, 2006.

Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Huiskes, M. J. and Lew, M. S. The mir flickr retrieval evaluation. In *ACM international conference on Multimedia information retrieval*, pp. 39–43, 2008.

Hwang, H., Kim, G.-H., Hong, S., and Kim, K.-E. Multi-view representation learning via total correlation objective. *Advances in Neural Information Processing Systems*, 34:12194–12207, 2021.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

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

Li, X. and Chen, S. A concise yet effective model for non-aligned incomplete multi-view and missing multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5918–5932, 2022.

Liu, C., Wu, Z., Wen, J., Xu, Y., and Huang, C. Localized sparse incomplete multi-view clustering. *IEEE Transactions on Multimedia*, pp. 5539–5551, 2022.

Liu, C., Wen, J., Luo, X., Huang, C., Wu, Z., and Xu, Y. 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, pp. 8807–8815, 2023a.

Liu, C., Wen, J., Luo, X., and Xu, Y. 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, pp. 8816–8824, 2023b.

Liu, C., Wen, J., Wu, Z., Luo, X., Huang, C., and Xu, Y. Information recovery-driven deep incomplete multiview clustering network. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–11, 2023c.

Liu, C., Wen, J., Liu, Y., Huang, C., Wu, Z., Luo, X., and Xu, Y. Masked two-channel decoupling framework for incomplete multi-view weak multi-label learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Liu, J., Liu, X., Yang, Y., Liao, Q., and Xia, Y. Contrastive multi-view kernel learning. 45(8):9552–9566, 2023d.

Liu, Y., Wang, J., Huang, C., Wang, Y., and Xu, Y. Cigar: Cross-modality graph reasoning for domain adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23776–23786, 2023e.

Liu, Y., Wang, J., Xiao, L., Liu, C., Wu, Z., and Xu, Y. Foregroundness-aware task disentanglement and self-paced curriculum learning for domain adaptive object detection. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–12, 2023f. doi: 10.1109/TNNLS.2023.3331778.

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

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

Ma, Q., Yuan, C., Zhou, W., and Hu, S. 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)*, pp. 3855–3864, 2021.

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

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

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

Wang, Q., Tao, Z., Gao, Q., and Jiao, L. Multi-view subspace clustering via structured multi-pathway network. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Wang, Q., Tao, Z., Xia, W., Gao, Q., Cao, X., and Jiao, L. Adversarial multiview clustering networks with adaptive fusion. *IEEE transactions on neural networks and learning systems*, 34:7635–7647, 2023.

Wen, J., Liu, C., Deng, S., Liu, Y., Fei, L., Yan, K., and Xu, Y. Deep double incomplete multi-view multi-label learning with incomplete labels and missing views. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Wu, M. and Goodman, N. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31, 2018.

Xiao, Y., Chen, J., Liu, B., Zhao, L., Kong, X., and Hao, Z. A new multi-view multi-label model with privileged information learning. *Information Sciences*, 656:119911, 2024.

Xu, J., Tang, H., Ren, Y., Peng, L., Zhu, X., and He, L. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pp. 16051–16060, 2022.

Yeh, C.-K., Wu, W.-C., Ko, W.-J., and Wang, Y.-C. F. Learning deep latent space for multi-label classification. In *AAAI Conference on Artificial Intelligence*, volume 31, 2017.

You, R., Guo, Z., Cui, L., Long, X., Bao, Y., and Wen, S. Cross-modality attention with semantic graph embedding for multi-label classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 12709–12716, 2020.

Zeng, P., Yang, M., Lu, Y., Zhang, C., Hu, P., and Peng, X. Semantic invariant multi-view clustering with fully incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

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

Zhao, D., Gao, Q., Lu, Y., and Sun, D. Non-aligned multi-view multi-label classification via learning view-specific labels. *IEEE Transactions on Multimedia*, 2022.

Zhu, Y., Kwok, J. T., and Zhou, Z.-H. Multi-label learning with global and local label correlation. *IEEE Transactions on Knowledge and Data Engineering*, 30(6):1081–1094, 2017.

A. Complete Derivation of Shared Information Learning Model

In this section, we give a detailed derivation of model (2):

$$\max \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}(I(\mathbf{x}^{(v)};\mathbf{z})-\beta I(\{\bar{\mathbf{x}}\};\mathbf{z}|\mathbf{x}^{(v)})) \tag{14}$$

For the former term in Model (2), we have:

$$\begin{aligned}
&I(\mathbf{x}^{(v)};\mathbf{z})\\
=&\int\int p(\mathbf{x}^{(v)},\mathbf{z})\log\frac{p(\mathbf{x}^{(v)}|\mathbf{z})}{p(\mathbf{x}^{(v)})}d\mathbf{x}^{(v)}d\mathbf{z}\\
=&\Big[\int\int p(\mathbf{x}^{(v)},\mathbf{z})\log p(\mathbf{x}^{(v)}|\mathbf{z})d\mathbf{x}^{(v)}d\mathbf{z}+\\
&\int p(\mathbf{z}|\mathbf{x}^{(v)})\int p(\mathbf{x}^{(v)})\log\frac{1}{p(\mathbf{x}^{(v)})}d\mathbf{x}^{(v)}d\mathbf{z}\Big]\\
=&\Big[\int\int p(\mathbf{x}^{(v)},\mathbf{z})\log p(\mathbf{x}^{(v)}|\mathbf{z})d\mathbf{x}^{(v)}d\mathbf{z}+H(\mathbf{x}^{(v)})\Big]
\end{aligned} \tag{15}$$

Due to the information entropy $H(x^{(v)})\geq 0$, we have

$$\begin{aligned}
&I(\mathbf{x}^{(v)};\mathbf{z})\\
\geq&\int\int p(\mathbf{x}^{(v)},\mathbf{z})\log p(\mathbf{x}^{(v)}|\mathbf{z})d\mathbf{x}^{(v)}d\mathbf{z}\\
=&\int p(\mathbf{x}^{(v)})\int p(\mathbf{z}|\mathbf{x}^{(v)})\log q^v(\mathbf{x}^{(v)}|\mathbf{z})d\mathbf{x}^{(v)}d\mathbf{z}+\\
&\int p(\mathbf{z})\int p(\mathbf{x}^{(v)}|\mathbf{z})\log\frac{p(\mathbf{x}^{(v)}|\mathbf{z})}{q^v(\mathbf{x}^{(v)}|\mathbf{z})}d\mathbf{x}^{(v)}d\mathbf{z}\\
=&\int p(\mathbf{x}^{(v)})\int p(\mathbf{z}|\mathbf{x}^{(v)})\log q^v(\mathbf{x}^{(v)}|\mathbf{z})d\mathbf{x}^{(v)}d\mathbf{z}+\\
&\int p(\mathbf{z})D_{KL}(p(\mathbf{x}^{(v)}|\mathbf{z})\|q^v(\mathbf{x}^{(v)}|\mathbf{z}))d\mathbf{x}^{(v)}d\mathbf{z}
\end{aligned} \tag{16}$$

Since $D_{KL}(p(\mathbf{x}^{(v)}|\mathbf{z})\|q(\mathbf{x}^{(v)}|\mathbf{z}))\geq 0$, we can get:

$$\begin{aligned}
&I(\mathbf{x}^{(v)};\mathbf{z})\\
\geq&\int p(\mathbf{x}^{(v)})\int p(\mathbf{z}|\mathbf{x}^{(v)})\log q^v(\mathbf{x}^{(v)}|\mathbf{z})d\mathbf{x}^{(v)}d\mathbf{z}\\
=&\int\int p(\mathbf{x}^{(v)},\mathbf{z})\log q^v(\mathbf{x}^{(v)}|\mathbf{z})d\mathbf{x}^{(v)}d\mathbf{z}
\end{aligned} \tag{17}$$

Besides, we have:

$$\begin{aligned}
\int\int p(\mathbf{x}^{(v)},\mathbf{z})d\mathbf{x}^{(v)}d\mathbf{z}&=\int\int\int p(\mathbf{x}^{(v)},\{\bar{\mathbf{x}}\},\mathbf{z})d\mathbf{x}^{(v)}d\{\bar{\mathbf{x}}\}d\mathbf{z}\\
&=\int\int p(\{\mathbf{x}\},\mathbf{z})d\{\mathbf{x}\}d\mathbf{z}
\end{aligned} \tag{18}$$

Substitute Eq. (18) into Eq. (17), we can get:

$$\begin{aligned}
&I(\mathbf{x}^{(v)};\mathbf{z})\\
\geq&\int\int p(\{\mathbf{x}\},\mathbf{z})\log q^v(\mathbf{x}^{(v)}|\mathbf{z})d\{\mathbf{x}\}d\mathbf{z}\\
=&\int p(\{\mathbf{x}\})\int p(\mathbf{z}|\{\mathbf{x}\})\log q^v(\mathbf{x}^{(v)}|\mathbf{z})d\{\mathbf{x}\}d\mathbf{z}\\
=&\mathbb{E}_{\mathbf{x}\sim p(\{\mathbf{x}\})}[\int p(\mathbf{z}|\{\mathbf{x}\})\log q^v(\mathbf{x}^{(v)}|\mathbf{z})d\mathbf{z}]
\end{aligned} \tag{19}$$

For the latter term in Model (2), we have:

$$\begin{aligned}
&I(\{\bar{\mathbf{x}}\};\mathbf{z}|\mathbf{x}^{(v)}) \\
=&\int\int p(\{\mathbf{x}\},\mathbf{z})\log\frac{p(\{\mathbf{x}\},\mathbf{z})p(\mathbf{x}^{(v)})}{p(\{\mathbf{x}\})p(\mathbf{z},\mathbf{x}^{(v)})}d\{\mathbf{x}\}d\mathbf{z} \\
=&\int\int p(\{\mathbf{x}\},\mathbf{z})\log\frac{p(\mathbf{z}|\{\{\mathbf{x}\}\})}{p(\mathbf{z}|\mathbf{x}^{(v)})}d\{\mathbf{x}\}d\mathbf{z} \\
=&\int\int p(\{\mathbf{x}\},\mathbf{z})\log\frac{p(\mathbf{z}|\{\mathbf{x}\})}{r^v(\mathbf{z}|\mathbf{x}^{(v)})}d\{\mathbf{x}\}d\mathbf{z}+ \\
&\int\int p(\{\mathbf{x}\},\mathbf{z})\log\frac{r^v(\mathbf{z}|\mathbf{x}^{(v)})}{p(\mathbf{z}|\mathbf{x}^{(v)})}d\{\mathbf{x}\}d\mathbf{z} \\
=&\int\int p(\{\mathbf{x}\},\mathbf{z})\log\frac{p(\mathbf{z}|\{\mathbf{x}\})}{r^v(\mathbf{z}|\mathbf{x}^{(v)})}d\{\mathbf{x}\}d\mathbf{z}+ \\
&\int p(\{\mathbf{x}\})\int p(\mathbf{z}|\{\mathbf{x}\})\log\frac{r^v(\mathbf{z}|\mathbf{x}^{(v)})}{p(\mathbf{z}|\mathbf{x}^{(v)})}d\{\mathbf{x}\}d\mathbf{z} \\
=&\int\int p(\{\mathbf{x}\},\mathbf{z})\log\frac{p(\mathbf{z}|\{\mathbf{x}\})}{r^v(\mathbf{z}|\mathbf{x}^{(v)})}d\{\mathbf{x}\}d\mathbf{z}+ \\
&\int p(\mathbf{x}^{(v)})\int p(\mathbf{z}|\mathbf{x}^{(v)})\log\frac{r^v(\mathbf{z}|\mathbf{x}^{(v)})}{p(\mathbf{z}|\mathbf{x}^{(v)})}d\mathbf{x}^{(v)}d\mathbf{z}
\end{aligned} \tag{20}$$

Since

$$\begin{aligned}
&\int p(\mathbf{x}^{(v)})\int p(\mathbf{z}|\mathbf{x}^{(v)})\log\frac{r^v(\mathbf{z}|\mathbf{x}^{(v)})}{p(\mathbf{z}|\mathbf{x}^{(v)})}d\mathbf{x}^{(v)}d\mathbf{z} \\
=&-\int p(\mathbf{x}^{(v)})D_{KL}(p(\mathbf{z}|\mathbf{x}^{(v)})\|r^v(\mathbf{z}|\mathbf{x}^{(v)}))d\mathbf{x}^{(v)}d\mathbf{z} \\
\leq&0
\end{aligned} \tag{21}$$

the Eq. (20) have following upper bound:

$$\begin{aligned}
&I(\{\bar{\mathbf{x}}\};\mathbf{z}|\mathbf{x}^{(v)}) \\
\leq&\int\int p(\{\mathbf{x}\},\mathbf{z})\log\frac{p(\mathbf{z}|\{\mathbf{x}\})}{r^v(\mathbf{z}|\mathbf{x}^{(v)})}d\{\mathbf{x}\}d\mathbf{z} \\
=&\int p(\{\mathbf{x}\})D_{KL}(p(\mathbf{z}|\{\mathbf{x}\})\|r^v(\mathbf{z}|\mathbf{x}^{(v)}))d\{\mathbf{x}\} \\
=&\mathbb{E}_{\{\mathbf{x}\}\sim p(\{\mathbf{x}\})}[D_{KL}(p(\mathbf{z}|\{\mathbf{x}\})\|r^v(\mathbf{z}|\mathbf{x}^{(v)}))]
\end{aligned} \tag{22}$$

By combining Eq. (19) and Eq. (22), we get the loss function corresponding to Model (2):

$$\begin{aligned}
\mathcal{L}_{IB}=\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\big[&-\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|\{\mathbf{x}\})}\log q^v(\mathbf{x}^{(v)}|\mathbf{z}) \\
&+\beta D_{KL}(p(\mathbf{z}|\{\mathbf{x}\})\|r^v(\mathbf{z}|\mathbf{x}^{(v)}))\big]
\end{aligned} \tag{23}$$

B. PoE Fusion and Sampling

Formulation of PoE fusion:

$$\begin{aligned}
\mu_{poe}&=\frac{\sum_{v\in\mathcal{V}}\mu_v\frac{1}{\sigma_v^2}}{\sum_{v\in\mathcal{V}}\frac{1}{\sigma_v^2}+1}, \\
\sigma_{poe}^2&=\frac{1}{\sum_{v\in\mathcal{V}}\frac{1}{\sigma_v^2}+1},
\end{aligned} \tag{24}$$

Table 3: Detailed information about five multi-view multi-label datasets in our experiments.

Dataset	# Sample	# Label	# View	# Label/#Sample
Corel5k	4999	260	6	3.40
IAPRTC12	19627	291	6	5.72
ESPGame	20770	268	6	4.69
Pascal07	9963	20	6	1.47
MIRFLICKR	25000	38	6	4.72

where μ_{poe} and σ^2_{poe} are the fused mean and variance of multiple views, respectively. μ_v and σ^2_v mean the v-th view's mean and variance, respectively. Then, we have $p(\mathbf{z}|\{\mathbf{x}\}) \sim \mathcal{N}(\mu_{poe}, \sigma^2_{poe}\mathbf{I})$.

Like Eq. (9), we sample z^0 from distribution $p(\mathbf{z}|\{\mathbf{x}\})$:

$$z^0 = \frac{1}{s}\sum_{d=1}^{s}(\mu_{poe} + \sigma_{poe} \odot \delta^d). \tag{25}$$

C. Statistics for Five Datasets

In this section, we give details of the five datasets used in the experiment in Table 3.

D. Statistics for Eight Competitors

In this section, we give details of the eight comparison methods in Table 4.

Table 4: Simple information of eight comparison methods. 'Multi-view' denotes the method is designed for multi-view data; 'Missing-view' and 'Missing-label' represent their compatibility with missing views and partial labels.

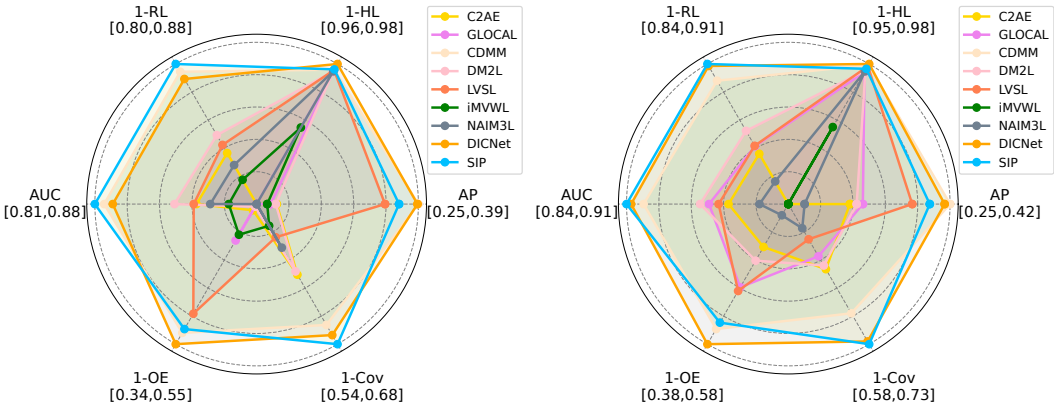
Method	Sources	Multi-view	Missing-view	Missing-label
C2AE	AAAI '17	✗	✗	✓
GLOCAL	TKDE '17	✗	✗	✓
CDMM	KBS '20	✓	✗	✗
DM2L	PR '21	✗	✗	✓
LVSL	TMM '22	✓	✗	✗
iMVWL	IJCAI '18	✓	✓	✓
NAIM3L	TPAMI '22	✓	✓	✓
DICNet	AAAI '23	✓	✓	✓

E. Extra Experimental Results on Five Full Datasets.

In this section, we show the results of nine methods on two datasets without any missing views and labels in Fig. 5.

F. Implementation Details

In the experiments, we set parameters as the values recommended in their codes or papers for all competitors. For our SIP, the batch size is 128, the dimension of latent feature is 512, learning rate is set as 0.001 for all five datasets and the Stochastic Gradient Descent (SGD) optimizer is employed to train the model. The sampling times are set as 10 for z^0 and l^0_i. Our encoder mainly consists of six Fully connected (FC) layers and the last three layers share parameters, while the decoder mainly consists of three FC layers. To avoid randomness, we repeat experiments for all methods multiple times and report the mean and variance in the final results. All the experiments are run on the ubuntu system with a 3090 GPU and pytorch 2.1.1 framework.



(a) ESPGame (b) MIRFLICKR

Figure 5: Experimental results of nine methods on the two full datasets without any missing views or labels. The worst results are indicated at the center of radar chart, while the best results are represented by the vertexes, considering six evaluation metrics.

G. Limitations

Although our method has demonstrated effectiveness in PMvMLC task, providing insights on utilizing shared information for prediction, there are limitations that restrict the broader application. A challenge is evaluating the assumption that shared information is sufficient for accurate predictions, especially when dealing with heterogeneous multi-view data with highly modality difference, such as text, images, and audio. Besides, many techniques have been developed for estimating information capacity, there remains potential to improve accuracy by employing more sophisticated estimators.