Theory-Inspired Deep Multi-View Multi-Label Learning with Incomplete Views and Noisy Labels

Quanjiang Li, Tingjin Luo*, Jiahui Liao
National University of Defense Technology, China
liquanjiang@nudt.edu.cn, tingjinluo@hotmail.com, liaojiahui19@nudt.edu.cn

Abstract

Incomplete features and label noise in multi-view multi-label data significantly undermine the reliability and performance, motivating researchers to explore the mechanism of representation and information recovery. However, learning for such dual deficiencies is crucial but rarely studied. In this paper, we propose a theory-inspired Deep Multi-View Multi-Label Learning method with Incomplete Views and Noisy Labels named DMMIvNL to address these problems. Specifically, to promote the synthesis of task-relevant shared information and preserve the distinctiveness of individual features from limited views, we have developed a feature extraction modular based on the information bottleneck theory, and formulated its theoretical upper bound into its objective. Meanwhile, we theoretically prove that minimizing the volume of the transition matrix ensures the statistical consistency with classifier training. Besides, a cycle-consistency estimation principle is proposed in the volume minimization network to improve the recognition stability of multi-label noise. Moreover, leveraging inherent real semantics information and label correlations are employed as model regularization to reduce the risk of excessive noise fitting. Finally, extensive experimental results validate the effectiveness and robustness of our DMMIvNL.

1. Introduction

Advances in feature extraction has shifted data analysis towards multi-view frameworks [27], while growing data attributes drive multi-label learning to offer rich label semantics [13]. Therefore, Multi-view multi-label classification (MvMLC) has emerged to address the intricate demands of processing diverse features and numerous targets. Traditional methods [30, 46] assume the given data has complete views and clean labels, which is violated in practice [41]. On one hand, heterogeneous data from diverse sources frequently encounters missing views due to acquisition complexities and storage limitations [29, 47]. On the other hand, real-world collections often contain label noise as manual annotation can introduce incorrect labels beyond human knowledge [34]. Therefore, insufficient features and noisy labels severely reduce the classification performance. Regarding MvMLC, many solutions have been proposed to address missing views. iMvWL [37] used weighted matrix factorization to learn a shared subspace and AIMNet [24] leveraged a cross-view attention mechanism to recover missing views. Moreover, two types of label deficiencies have been considered in MvMLC. The problem of Partial Multi-Label (PML) [40] involves both relevant and irrelevant labels, while the problem of Multi-Label Learning with Missing Labels (MLML)[36] occurs when some relevant labels are missing. TFMDD [38] addressed PML with a two-stage manifold disambiguation model and LMVCAT [23] tackled MLML by employing label-guided graphs and category-aware transformers. While existing methods address each annotation problem separately, both issues tend to occur simultaneously in practice. Thus, the challenge of multi-label noise demands substantial attention, where each category suffers from both over-labeled and missing labels.

Existing research has not yet tackled the combined issues of missing views and multi-label noise, which are prevalent in real-world applications like incomplete and subjective medical diagnostics [31]. In fact, overcoming the simultaneous challenges requires refined representation extraction under missing views and effective classifier leaning with multi-label noise. The core of multi-view learning lies in extracting shared information across different views while preserving unique features of each view [33]. DICNet [22], DIMC [39] and SIP [26] concentrated solely on shared subspace and neglected distinctive characteristics of individual views. MDDRL [16] and MTD [25] attempted to extract both shared and specific representations with minimal interactions. However, their reliance on geometric orthogonality or distance metrics restricted them to capture only linear patterns, overlooking the complex nature of inter-view relations. Therefore, constraining feature semantics at the intrinsic information level is crucial for obtaining high-level

*Corresponding author.

representations with precise distinction. Most approaches for handling multi-label noise utilize strategies like early stopping [14] and cyclical learning rates [3]. These methods often lack stability and require dataset-specific adjustments, making them insufficiently adaptive. ENTMLC [19] employed label correlations to identify the noise transition matrix for recovering clean labels. However, the decoupling of transition matrix acquisition and classifier training causes error accumulation, limiting the practical utility. To circumvent the uncertainty inherent in those strategies and the inefficiency of stepwise noise learning, a statistically consistent estimation framework should be developed, which would allow for direct training the optimal classifier from noisy labels and achieving strong robustness.

To tackle these problems, we propose a theory-inspired method with deep networks named DMMIvNL. Specifically, a mutual information model is developed to enhance synergy of shared features, suppress interactions from specific features, and impose orthogonality between shared and specific features of each view. The upper bound of the model is derived through the variational autoencoder framework, with instance-level contrastive operations applied to simulate diverse joint and marginal distributions. Moreover, the ideal classifier is trained on noisy data by minimizing the volume of the noise transition matrix and the process is proven to converge to a unique and globally optimal solution under the sufficiently scattered assumption. The difficulty in guaranteeing the assumption and relying exclusively on noisy class-posterior estimation promotes the design of a cycle-consistency estimation method, which integrates the volume maximization of clean class-posterior. Finally, the real semantics embedded in noisy data and hidden label correlation information are employed as model regularization to reduce the risk of noise overfitting. Our main contributions are summarized as follows:

- This is the first multi-view multi-label learning framework capable of handling incomplete views and noisy labels, including both over-labeled and missing labels.
- Our mutual information based model extracts semantically distinct representations from limited views, with its upper bound established through variational derivation.
- We theoretically prove the statistical consistency and joint convergence between minimizing the noise transition matrix volume for each category and classifier training. Besides, a cycle-consistency loss framework is designed to enhance transition matrix estimation.

2. Methodology

2.1. Problem Definition

The multi-view dataset $\{\boldsymbol{X}, \boldsymbol{Y}\}$ contains N samples, where $\boldsymbol{X}_i$ consists of V views with dimension d_v, i,e, $\boldsymbol{X}_i = \{\boldsymbol{x}_i^{(v)} \in \mathbb{R}^{d_v}\}_{v=1}^V$. The label matrix $\boldsymbol{Y} \in \{0,1\}^{N \times C}$ corresponds to C categories, with $Y_{i,j} = 1$ if sample i belongs to class j, and $Y_{i,j} = 0$ otherwise. $\boldsymbol{W} \in \mathbb{R}^{N \times V}$ is an indicator matrix, where $W_{i,j} = 1$ indicates the j-th view is available for sample i, and $W_{i,j} = 0$ means missing. Since precise annotation is hard to obtain, the training dataset in practice is $\{\boldsymbol{X}, \bar{\boldsymbol{Y}}\}$, where $\bar{\boldsymbol{Y}}$ denotes noisy labels. Thus, we denote the clean and noisy labels of the j-th class as $\boldsymbol{Y}^j$ and $\bar{\boldsymbol{Y}}^j$, related by a noise transition matrix $\boldsymbol{T}^j \in [0,1]^{2\times 2}, j \in [C]$. When only noisy samples are available and without further constraints, instance-dependent transition matrix becomes non-identifiable [42]. Therefore, we employ a general assumption that the noise transition matrix is class-dependent and instance-independent, i.e., $P(\bar{\boldsymbol{Y}}^j = k \mid \boldsymbol{Y}^j = i, \boldsymbol{X} = \boldsymbol{x}) = P(\bar{\boldsymbol{Y}}^j = k \mid \boldsymbol{Y}^j = i) = T^j_{i,k}$.

2.2. Feature Extraction via Information Bottleneck

Deriving effective representations from raw data aids in reducing noise and redundancy [17]. Consequently, both task-relevant shared semantics and view-invariant characteristic features should be explored to enhance information processing efficiency. Let $\boldsymbol{z}^v$ and $\boldsymbol{s}^v$ denote the D-dimensional shared and specific representation extracted from each view and the goal is to ensure consistency of the shared features across multiple views and preserve the unique contributions of each view. Considering different interaction patterns between representations from $\boldsymbol{x}^{v^*}$ and $\boldsymbol{x}^v$, mutual information $I(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*}; \boldsymbol{z}^v|\boldsymbol{x}^v)$ should be maximized to capture only universally shared information, while $I(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*}; \boldsymbol{s}^v|\boldsymbol{x}^v)$ should be minimized to preserve only individual attributes. Besides, the shared and specific features should be as orthogonal as possible to ensure absolute distinction in information content [35], i.e, $I(\boldsymbol{z}^v|\boldsymbol{x}^v; \boldsymbol{s}^v|\boldsymbol{x}^v) = 0$. Thus, the constrained optimization problem based on the information bottleneck theory [8] can be formulated as

$$\begin{aligned} &\max \frac{1}{V}\sum_{v=1}^{V}\sum_{v^*\neq v}^{V}\left(I(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*}; \boldsymbol{z}^v|\boldsymbol{x}^v) - I(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*}; \boldsymbol{s}^v|\boldsymbol{x}^v)\right) \\ &s.t. \quad \frac{1}{V}\sum_{v=1}^{V} I(\boldsymbol{z}^v|\boldsymbol{x}^v; \boldsymbol{s}^v|\boldsymbol{x}^v) = 0. \end{aligned} \tag{1}$$

By applying the Lagrange multiplier method, the problem (1) can be equally converted as

$$\begin{aligned} &\max \frac{1}{V}\sum_{v=1}^{V}\sum_{v^*\neq v}^{V}\left(I(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*}; \boldsymbol{z}^v|\boldsymbol{x}^v) - I(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*}; \boldsymbol{s}^v|\boldsymbol{x}^v)\right) \\ &- \frac{\beta}{V}\sum_{v=1}^{V} I(\boldsymbol{z}^v|\boldsymbol{x}^v; \boldsymbol{s}^v|\boldsymbol{x}^v), \end{aligned} \tag{2}$$

where $\beta \geq 0$ is the Lagrange multiplier to balance the efficiency and compactness of information [26]. For the first

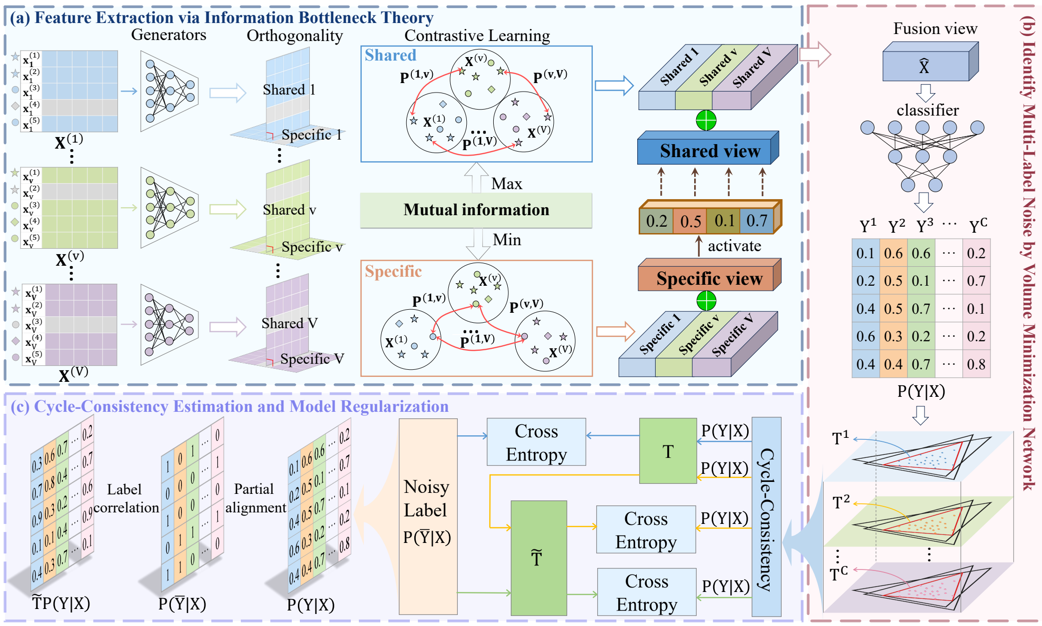


Figure 1. The main framework of our proposed DMMIvNL. Different shapes signify different samples.

term, we can obtain the following lower bound:

$$
\begin{aligned}
& I(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{z}^v|\boldsymbol{x}^v) \\
&= \int\int p(\boldsymbol{z}^v|\boldsymbol{x}^v,\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*}) \log \frac{p(\boldsymbol{z}^v|\boldsymbol{x}^v,\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})}{p(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*})p(\boldsymbol{z}^v|\boldsymbol{x}^v)} d\boldsymbol{z}^{v^*} d\boldsymbol{z}^v \\
&\geq \int\int p(\boldsymbol{z}^v|\boldsymbol{z}^{v^*}) p(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*}) \log p(\boldsymbol{z}^v|\boldsymbol{z}^{v^*}) d\boldsymbol{z}^{v^*} d\boldsymbol{z}^v,
\end{aligned}
$$

where $p(\cdot)$ denotes the probability density function. Since computing $p(\boldsymbol{z}^v|\boldsymbol{z}^{v^*})$ is intractable, we approximate it using a stochastic variational distribution $q(\boldsymbol{z}^v|\boldsymbol{z}^{v^*})$. Then we can obtain a new variational lower bound for maximization:

$$
\begin{aligned}
& I(\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{z}^v|\boldsymbol{x}^v) \\
&\geq \int\int p(\boldsymbol{z}^v|\boldsymbol{x}^v,\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*}) \log q(\boldsymbol{z}^v|\boldsymbol{z}^{v^*}) d\boldsymbol{z}^{v^*} d\boldsymbol{z}^v \\
&= D^{\dagger}_{KL}(p(\boldsymbol{z}^v|\boldsymbol{x}^v,\boldsymbol{z}^{v^*}|\boldsymbol{x}^{v^*}) \| q(\boldsymbol{z}^v|\boldsymbol{z}^{v^*})) = D^{\dagger(\boldsymbol{z}^v,\boldsymbol{z}^{v^*})}_{KL},
\end{aligned} \tag{3}
$$

where $D_{KL}(\cdot||\cdot)$ means the Kullback-Leibler divergence and $D^{\dagger}_{KL}(\cdot||\cdot)$ is a pseudo-symbol for notational consistency. For the middle term, the upper bound can be obtained by introducing another variational distribution $q(\boldsymbol{s}^v|\boldsymbol{s}^{v^*})$:

$$
\begin{aligned}
& I(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*};\boldsymbol{s}^v|\boldsymbol{x}^v) \\
&\leq D_{KL}(p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*},\boldsymbol{s}^v|\boldsymbol{x}^v) \| p(\boldsymbol{s}^{v^*}|\boldsymbol{x}^{v^*}) q(\boldsymbol{s}^v|\boldsymbol{s}^{v^*})) \\
&= D^{(\boldsymbol{s}^v,\boldsymbol{s}^{v^*})}_{KL}.
\end{aligned} \tag{4}
$$

Similarly, the upper bound for the last term is

$$
I(\boldsymbol{z}^v|\boldsymbol{x}^v;\boldsymbol{s}^v|\boldsymbol{x}^v) \leq D^{(\boldsymbol{s}^v,\boldsymbol{z}^v)}_{KL}. \tag{5}
$$

Combining Eqs. (3), (4) and (5), our optimization objective becomes minimizing the upper bound of the problem (2):

$$
\mathcal{L}_{IB} = \frac{1}{V}\sum_{v=1}^{V}\sum_{v^*\neq v}^{V}(-D^{\dagger(\boldsymbol{z}^v,\boldsymbol{z}^{v^*})}_{KL} + D^{(\boldsymbol{s}^v,\boldsymbol{s}^{v^*})}_{KL}) + \frac{\beta}{V}\sum_{v=1}^{V} D^{(\boldsymbol{s}^v,\boldsymbol{z}^v)}_{KL}. \tag{6}
$$

Contrastive learning [21] is used to compute various distributions. Specifically, representations are treated as probability vectors over D classes via a Softmax activation function, and the joint distribution matrix is obtained by

$$
\boldsymbol{P}^{(v,v^*)} = \sum_{i=1}^{N}\left(\boldsymbol{z}_i^{(v)}\right)^T \boldsymbol{z}_i^{(v^*)}, \boldsymbol{Q}^{(v,v^*)} = \sum_{i=1}^{N}\left(\boldsymbol{s}_i^{(v)}\right)^T \boldsymbol{s}_i^{(v^*)}.
$$

Due to the strong interaction between $\boldsymbol{z}^v$ and $\boldsymbol{z}^{v^*}$, the variational distribution $q(\boldsymbol{z}^v|\boldsymbol{z}^{v^*})$ can be estimated by the obtained conditional distribution. Given the extremely weak correlation between $\boldsymbol{s}^v$ and $\boldsymbol{s}^{v^*}$, as well as $\boldsymbol{z}^v$ and $\boldsymbol{s}^v$, the distribution $q(\boldsymbol{s}^v/\boldsymbol{s}^{v^*})$ and $q(\boldsymbol{z}^v/\boldsymbol{s}^v)$ can be approximated by the corresponding marginal distribution. Subsequently, each term in Eq. (6) can be expressed as

$$
\begin{cases}
D^{\dagger(\boldsymbol{z}^v,\boldsymbol{z}^{v^*})}_{KL} = \sum_{d=1}^{D}\sum_{d'=1}^{D} \boldsymbol{P}^{(v,v^*)}_{d,d'} \log \frac{\boldsymbol{P}^{(v,v^*)}_{d,d'}}{\boldsymbol{P}^{(v)}_{d}} \\
D^{(\boldsymbol{s}^v,\boldsymbol{s}^{v^*})}_{KL} = \sum_{d=1}^{D}\sum_{d'=1}^{D} \boldsymbol{Q}^{(v,v^*)}_{d,d'} \log \frac{\boldsymbol{Q}^{(v,v^*)}_{d,d'}}{(\boldsymbol{Q}^{(v)}_{d})^{\alpha}\boldsymbol{Q}^{(v^*)}_{d'}} \\
D^{(\boldsymbol{s}^v,\boldsymbol{z}^v)}_{KL} = \sum_{d=1}^{D}\sum_{d'=1}^{D} \boldsymbol{M}^{(\boldsymbol{z}^v,\boldsymbol{s}^v)}_{d,d'} \log \frac{\boldsymbol{M}^{(\boldsymbol{z}^v,\boldsymbol{s}^v)}_{d,d'}}{(\boldsymbol{M}^{(\boldsymbol{z}^v)}_{d})^{\alpha}\boldsymbol{M}^{(\boldsymbol{s}^v)}_{d'}},
\end{cases} \tag{7}
$$

where the marginal distributions $\boldsymbol{P}_d^{(v)}$ and $\boldsymbol{P}_{d'}^{(v^*)}$ can be obtained by summing over the d-th row and the d'-th column of the joint distribution matrix $\boldsymbol{P}^{(v,v^*)}$, with other symbols defined as the same. In this conversion, we let $q(\boldsymbol{s}^v/\boldsymbol{s}^{v^*}) = (\boldsymbol{Q}^{(v)})^\alpha$ and $q(\boldsymbol{z}^v/\boldsymbol{s}^v) = (\boldsymbol{M}^{(z^v)})^\alpha$, where α is a balance factor to preserve crucial information and ensure model stability [16]. By introducing the mask matrix $\boldsymbol{W}$ to mitigate the impact of unavailable instances, we can perform cross-view fusion to acquire the information-enriched shared and specific representation:

$$\bar{\boldsymbol{Z}}_{i,:} = \sum_{v=1}^{V} \frac{\boldsymbol{z}_{i,:}^{(v)} \boldsymbol{W}_{i,v}}{\sum_v \boldsymbol{W}_{i,v}}, \quad \bar{\boldsymbol{S}}_{i,:} = \sum_{v=1}^{V} \frac{\boldsymbol{s}_{i,:}^{(v)} \boldsymbol{W}_{i,v}}{\sum_v \boldsymbol{W}_{i,v}}. \tag{8}$$

Departing from traditional addition or concatenation, we conduct feature interaction by enhancing shared information with proprietary data [25]:

$$\hat{\boldsymbol{X}}_{i,j} = \theta_s(\bar{\boldsymbol{Z}}_{i,j}) \cdot \bar{\boldsymbol{S}}_{i,j}, \tag{9}$$

where θ_s is the sigmoid activation function and $\hat{\boldsymbol{X}} \in \mathbb{R}^{N \times D}$ is the final fusion representation.

2.3. Identify Multi-Label Noise by Volume Minimization Network

The success of deep learning relies on large-scale accurately labeled data, with performance significantly dropping under noisy data [12], making noise identification crucial for network training [2]. In multi-class tasks, many statistically consistent methods based on the transition matrix have been developed [28, 32], which guarantees the classifier learned from noisy data can converge to the optimal classifier under the clean risk. The anchor-point assumption [28] is commonly utilized to determine the transition matrix, which postulates the existence of samples that belong to each class with probability one. However, implementing this approach in multi-label learning is challenging due to the difficulty of pinpointing suitable anchor samples for each label. Motivated by [20], we capitalize on the geometric properties of the transition matrix to seek a breakthrough. The transition matrix for the j-th label is defined as

$$\boldsymbol{T}^j = \begin{pmatrix} P(\bar{\boldsymbol{Y}}^j = 0 \mid \boldsymbol{Y}^j = 0) & P(\bar{\boldsymbol{Y}}^j = 0 \mid \boldsymbol{Y}^j = 1) \\ P(\bar{\boldsymbol{Y}}^j = 1 \mid \boldsymbol{Y}^j = 0) & P(\bar{\boldsymbol{Y}}^j = 1 \mid \boldsymbol{Y}^j = 1) \end{pmatrix}. \tag{10}$$

Given an instance $\hat{\boldsymbol{x}}$, the noisy class-posterior for the j-th label is $P(\bar{\boldsymbol{Y}}^j|\boldsymbol{X} = \hat{\boldsymbol{x}}) = [P(\bar{\boldsymbol{Y}}^j = 0|\boldsymbol{X} = \hat{\boldsymbol{x}}), P(\bar{\boldsymbol{Y}}^j = 1|\boldsymbol{X} = \hat{\boldsymbol{x}})]^T$ and the clean class-posterior is $P(\boldsymbol{Y}^j|\boldsymbol{X} = \hat{\boldsymbol{x}}) = [P(\boldsymbol{Y}^j = 0|\boldsymbol{X} = \hat{\boldsymbol{x}}), P(\boldsymbol{Y}^j = 1|\boldsymbol{X} = \hat{\boldsymbol{x}})]^T$. Thus, the relationship $P(\bar{\boldsymbol{Y}}^j|\boldsymbol{X} = \hat{\boldsymbol{x}}) = \boldsymbol{T}^j P(\boldsymbol{Y}^j|\boldsymbol{X} = \hat{\boldsymbol{x}})$ holds. Since $\sum_{i=0}^{1} P(\boldsymbol{Y} = i|\boldsymbol{X} = \hat{\boldsymbol{x}}) = 1$, $P(\bar{\boldsymbol{Y}}^j|\boldsymbol{X} = \hat{\boldsymbol{x}})$ is then a convex combination of the columns of $\boldsymbol{T}^j$, which implies that the noisy class-posterior is enclosed within the simplex $\mathrm{Sim}\{\boldsymbol{T}^j\}$. To reduce noise identification errors, it is essential to recover $\mathrm{Sim}\{\boldsymbol{T}^j\}$ such that it tightly encompasses all sample posterior probability points, thereby making volume minimization of $\mathrm{Sim}\{\boldsymbol{T}^j\}$ the primary goal. For synchronized learning, we define a trainable $\hat{\boldsymbol{T}}^j \in [0,1]^{2\times 2}$ with $\sum_{i=1}^{2} \hat{\boldsymbol{T}}^j{}_{i,k} = 1$, and construct a multi-label classifier h_θ with parameters θ. The optimization of $\hat{\boldsymbol{T}}^j$ and $h_\theta(\hat{\boldsymbol{x}})$ involves two objectives: the discrepancy between $\hat{\boldsymbol{T}}^j(h_\theta(\hat{\boldsymbol{x}})_j, 1 - h_\theta(\hat{\boldsymbol{x}})_j)^T$ and $P(\bar{\boldsymbol{Y}}^j|\boldsymbol{X} = \hat{\boldsymbol{x}})$ and the volume of each simplex spanned by the columns of $\hat{\boldsymbol{T}}^j$. Therefore, the following model to jointly learn the transition matrix and classifier is proposed:

$$\begin{cases} \min\limits_{\hat{\boldsymbol{T}}^j} \sum\limits_{j=1}^{C} \mathrm{vol}(\hat{\boldsymbol{T}}^j) \\ s.t.\ \hat{\boldsymbol{T}}^j \begin{bmatrix} h_\theta(\hat{\boldsymbol{x}})_j \\ 1 - h_\theta(\hat{\boldsymbol{x}})_j \end{bmatrix} = P(\bar{\boldsymbol{Y}}^j|X = \hat{\boldsymbol{x}}), \end{cases} \tag{11}$$

where $\mathrm{vol}(\hat{\boldsymbol{T}}^j)$ denotes a measure to the volume of $\mathrm{Sim}\{\hat{\boldsymbol{T}}^j\}$. In the absence of anchor points, the class-posterior probability should incorporate substantial information to uniquely determine the minimum-volume simplex, with true sample probability points of each class being distinguishable. Therefore, the sufficiently scattered assumption [9] is introduced to limit the occurrence of anomalous clean class-posterior and aid in theoretically establishing that addressing problem (11) will drive $\hat{\boldsymbol{T}}^j$ to converge to $\boldsymbol{T}^j$ and $h_\theta(\hat{\boldsymbol{x}})$ to accurately approximate $P(\boldsymbol{Y}|\boldsymbol{X} = \hat{\boldsymbol{x}})$:

Theorem 1. *If each $P(\boldsymbol{Y}^j|\boldsymbol{X})$ is sufficiently scattered, then $\boldsymbol{T}_\star^j = \boldsymbol{T}^j$ and $(h_{\theta_\star}(\hat{\boldsymbol{x}})_j, 1 - h_{\theta_\star}(\hat{\boldsymbol{x}})_j)^T = P(\boldsymbol{Y}^j|\boldsymbol{X} = \hat{\boldsymbol{x}})$ $(j = 1, 2 \cdots C)$ must hold, where $(\boldsymbol{T}_\star^1, \boldsymbol{T}_\star^2, \cdots, \boldsymbol{T}_\star^C, \theta_\star)$ are the optimal solutions of the problem (11).*

2.4. Cycle-Consistency Estimation and Model Regularization

In multi-label learning, imbalanced label distributions [44] and inter-class correlations [7] can compress class-posterior probability points and limit their scatter. Besides, the volume approximation of $\boldsymbol{T}^j$ in the problem (11) is highly sensitive to the noisy class-posterior, making it suffer from large estimation errors in noisy environments [18]. The failure of the sufficiently scattered assumption, along with the excessive reliance on the noisy class-posterior, compromises the effectiveness of the problem (11). To enhance transition matrix estimation, we develop an adaptive loss framework [4] that incorporates the volume maximization of the clean class-posterior, which leverages additional information to mitigate bias in single-target decision-making. Based on the equation $P(\bar{\boldsymbol{Y}}^j|\boldsymbol{X}) = \boldsymbol{T}^j P(\boldsymbol{Y}^j|\boldsymbol{X})$, where $P(\bar{\boldsymbol{Y}}^j|\boldsymbol{X})$ can be derived from the noisy data, minimizing the volume of $\boldsymbol{T}^j$ is equivalent to imposing a volume constraint on $P(\boldsymbol{Y}^j|\boldsymbol{X})$. Through the matrix $\tilde{\boldsymbol{T}}^j$ (an approximate inverse of $\boldsymbol{T}^j$), we project the simplex formed in the noisy space onto the clean class-posterior space to obtain $P'(\boldsymbol{Y}|\boldsymbol{X})$, which will make up the largest volume. Then we design a cycle-consistency optimization strategy to enhance both the volume constraint on $\boldsymbol{T}^j$ and maximization of $P(\boldsymbol{Y}|\boldsymbol{X})$, in which the estimated noisy class-posterior matches the provided noisy label $\bar{\boldsymbol{Y}}^j$, the backward estimated clean class-posterior $P'(\boldsymbol{Y}|\boldsymbol{X})$ aligns with the predicted clean-class posterior and the forward-backward consistency term $\tilde{\boldsymbol{T}}^j\boldsymbol{T}^j P(\boldsymbol{Y}|\boldsymbol{X})$ coheres with $P(\boldsymbol{Y}|\boldsymbol{X})$.

Firstly, we employ forward estimation to simultaneously optimize the transition matrix $\boldsymbol{T}^j$ and classifier h_θ. By leveraging the cross-entropy loss and reporting the output as $\boldsymbol{F}^{ij} = (h_\theta(\hat{\boldsymbol{x}}_i)_j, 1-h_\theta(\hat{\boldsymbol{x}}_i)_j)^T$, we aim to minimize the approximation error between the inferred noisy class-posterior probability $\boldsymbol{T}^j\boldsymbol{F}^{ij}$ and the given noisy label $\bar{\boldsymbol{Y}}^j$:

$$\begin{aligned}\mathcal{L}_1(\theta, \boldsymbol{T}^1, \cdots, \boldsymbol{T}^C) = &-\frac{1}{CN}\sum_{i=1}^{N}\sum_{j=1}^{C}(\bar{Y}_{i,j}\log(\boldsymbol{T}_1^j\boldsymbol{F}^{ij})\\ &+(1-\bar{Y}_{i,j})\log(1-\boldsymbol{T}_2^j\boldsymbol{F}^{ij})).\end{aligned}$$

Secondly, the Sigmoid function is employed to convert the one-hot label into the noisy class-posterior probability, which is then normalized along the label dimension to capture the soft labels $\bar{\boldsymbol{Y}}^{sn}$. Denoting the outcome $(\bar{\boldsymbol{Y}}_{i,j}^{sn}, 1-\bar{\boldsymbol{Y}}_{i,j}^{sn})^T$ as $\bar{\boldsymbol{F}}^{ij}$, we utilize $\tilde{\boldsymbol{T}}^j$ to project $\bar{\boldsymbol{F}}^{ij}$ onto the clean class-posterior space. Thus, we minimize the following cross-entropy loss between the backward estimated clean class-posterior $\tilde{\boldsymbol{T}}^j\bar{\boldsymbol{F}}^{ij}$ and the true class-posterior probability estimation $h_\theta(\hat{\boldsymbol{x}}_i)_j$:

$$\begin{aligned}\mathcal{L}_2(\theta, \tilde{\boldsymbol{T}}^1, \cdots, \tilde{\boldsymbol{T}}^C) = &-\frac{1}{CN}\sum_{i=1}^{N}\sum_{j=1}^{C}(h_\theta(\hat{\boldsymbol{x}}_i)_j\log(\tilde{\boldsymbol{T}}_1^j\bar{\boldsymbol{F}}^{ij})\\ &+(1-h_\theta(\hat{\boldsymbol{x}}_i)_j)\log(1-\tilde{\boldsymbol{T}}_2^j\bar{\boldsymbol{F}}^{ij})).\end{aligned}$$

At this stage, the transition matrix can be derived from both the forward and backward ways by optimizing $\mathcal{L}_1$ and $\mathcal{L}_2$. Thirdly, we should establish the cycle-consistency loss by minimizing the difference between $\tilde{\boldsymbol{T}}^j\boldsymbol{T}^j\boldsymbol{F}^{ij}$ and $h_\theta(\hat{\boldsymbol{x}}_i)_j$ to reinforce the invertible relationship of $\boldsymbol{T}^j$ and $\tilde{\boldsymbol{T}}^j$:

$$\begin{aligned}\mathcal{L}_3(\theta) = &-\frac{1}{CN}\sum_{i=1}^{N}\sum_{j=1}^{C}(h_\theta(\hat{\boldsymbol{x}}_i)_j\log(\tilde{\boldsymbol{T}}_1^j\boldsymbol{T}^j\boldsymbol{F}^{ij})\\ &+(1-h_\theta(\hat{\boldsymbol{x}}_i)_j)\log(1-\tilde{\boldsymbol{T}}_2^j\boldsymbol{T}^j\boldsymbol{F}^{ij})),\end{aligned} \quad (12)$$

where we only optimize the network parameters θ while keeping $\boldsymbol{T}^j$ and $\tilde{\boldsymbol{T}}^j$ constant in order to maintain their consistency and focus on learning an optimal classifier. Therefore, the loss for identifying the transition matrix and learning the clean class-posterior is $\mathcal{L}_{IB} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3$.

To mitigate overfitting on noisy data, we design model regularization by leveraging the partial fidelity of noisy labels and label correlations. We allow the classifier to directly access raw labels and obtain the classification loss:

$$\begin{aligned}\mathcal{L}_c = &-\frac{1}{CN}\sum_{i=1}^{N}\sum_{j=1}^{C}(\bar{Y}_{i,j}\log(h_\theta(\hat{\boldsymbol{x}}_i)_j))\\ &+(1-\bar{Y}_{i,j})\log(1-h_\theta(\hat{\boldsymbol{x}}_i)_j)).\end{aligned} \quad (13)$$

The construction of $\mathcal{L}_c$ is motivated by two key considerations. On one hand, it is essential to harness the underlying true information embedded in the noisy data to improve the inefficient classifier. On the other hand, strengthening the model memory with original noise helps minimize the risk of misclassifying clean samples as noisy ones. Since label correlation is a crucial factor for multi-label learning, we minimize the discrepancy in the label relationships reflected by the estimated clean class posterior and the classifier output to reduce classification bias. By defining the clean class-posterior matrix of the positive labels as $\boldsymbol{P}^c$, with elements given by $\boldsymbol{P}_{i,j}^c = \tilde{\boldsymbol{T}}_1^j\bar{\boldsymbol{F}}^{ij}$, we can build the following correlation-aligned loss driven by label co-occurrence [43]:

$$\mathcal{L}_r = \frac{1}{C^2}\left\|(\boldsymbol{P}^c)^T\boldsymbol{P}^c - \left(h_\theta(\hat{\boldsymbol{X}})\right)^T h_\theta(\hat{\boldsymbol{X}})\right\|_F^2. \quad (14)$$

Thus, our model regularization can be formulated as $\mathcal{L}_R = \mathcal{L}_c + \mathcal{L}_r$. The total training loss of our model is

$$\mathcal{L} = \mathcal{L}_{IB} + \lambda_1\mathcal{L}_{ID} + \lambda_2\mathcal{L}_R, \quad (15)$$

where λ_1 and λ_2 are penalty coefficients to balance the effects of different losses when handling noisy labels.

3. Experiments

3.1. Experimental Setting

Datasets and Evaluation Metrics. Following [23, 45], our experiments incorporate six widely recognized multi-view multi-label datasets, i.e, Yeast [11], Corel 5k [5], VOC 2007 [6], Espgame [1], IAPR TC-12 [10], and MIR FLICKR [15]. Similar to [25], we employ six diverse metrics to comprehensively evaluate each method, i.e., Hamming Loss (HL), Ranking Loss (RL), OneError (OE), Coverage (Cov), Average Precision (AP), and Area Under Curve (AUC). To facilitate performance comparison, we present 1-HL, 1-RL, 1-OE, and 1-Cov values in our report, with higher values consistently indicating better performance.

Comparison Methods. To validate the effectiveness of DMMIvNL, we perform a comparative analysis against eight state-of-the-art methods i.e., iMVWL, TM3L, AIMNet, DICNet, DIMC, MTD, SIP, ENTMLC. Since the first seven methods are designed to handle incomplete multi-view multi-label problems, we train them directly on noisy data without modifications. As ENTMLC specializes in addressing multi-label noise, we tackle missing views by fill-

Table 1. Experimental results of nine methods on all datasets with PER=50% and $\rho = 30\%$. The best result on each row is bolded and the second-best result is underlined. The average ranking on the six metrics is shown as Ave.R.

Data	MET	DICNet	DIMC	MTD	SIP	AIMNet	iMVWL	TM3L	NTMLCS	DMMIvNL
Yeast	1-HL	71.54±1.81	74.94±1.56	76.40±0.33	71.24±1.49	73.03±1.79	66.94±0.47	63.89±0.29	69.88±0.16	**77.64±0.47**
	1-RL	73.81±0.91	74.82±2.33	77.08±0.75	71.30±1.67	73.95±0.43	72.83±1.05	65.72±0.37	71.40±3.14	**80.18±0.37**
	1-OE	71.86±3.74	70.95±5.37	**75.33±0.48**	61.94±6.74	70.66±2.79	63.91±2.29	56.49±0.77	73.28±0.43	73.69±1.22
	1-Cov	41.68±2.48	44.12±3.10	45.53±1.95	39.04±2.16	41.19±0.31	41.50±1.45	34.93±0.73	35.55±1.37	**50.77±0.84**
	AP	65.61±1.34	66.91±2.09	68.87±0.82	61.78±2.25	65.55±1.18	63.74±0.92	56.90±0.30	66.69±0.60	**72.08±0.34**
	AUC	75.33±0.73	76.34±2.35	78.36±0.54	73.42±1.53	75.46±0.45	74.85±0.97	67.92±0.28	73.33±2.44	**81.03±0.55**
	Ave.R	4.50	3.33	1.83	7.17	4.83	6.33	8.83	7.00	**1.17**
Corel 5k	1-HL	96.37±0.68	96.08±0.99	98.65±0.00	98.59±0.04	98.66±0.00	79.41±0.05	97.23±0.67	98.57±0.17	**98.68±0.02**
	1-RL	60.14±0.84	61.00±0.87	67.26±0.59	65.32±1.41	63.37±0.65	64.53±0.55	52.01±0.39	65.92±1.23	**72.27±0.57**
	1-OE	6.30±0.89	5.44±1.05	21.22±1.06	12.88±4.28	24.00±0.73	10.29±2.46	20.18±0.33	2.30±1.67	**25.20±0.78**
	1-Cov	32.29±0.81	33.31±0.96	37.53±0.88	36.69±2.03	41.62±0.47	36.09±1.12	33.03±0.02	35.19±0.96	**44.35±1.10**
	AP	6.72±0.40	6.80±0.29	15.22±0.65	10.54±1.86	18.49±0.21	10.18±0.91	4.48±0.09	8.25±0.99	**19.18±0.45**
	AUC	60.30±0.85	51.74±1.30	67.63±0.59	65.69±1.54	70.37±0.37	64.81±0.76	58.99±0.11	65.97±1.27	**72.42±0.47**
	Ave.R	7.67	7.67	2.83	4.33	2.67	6.00	7.33	5.50	**1.00**
VOC 2007	1-HL	89.48±1.70	89.72±1.31	**92.85±0.05**	92.00±0.17	92.23±0.32	74.95±0.20	84.92±0.20	92.66±0.02	92.67±0.06
	1-RL	64.72±0.89	65.70±0.76	66.04±1.10	63.69±1.72	63.56±0.92	67.03±0.81	56.30±0.58	66.03±1.95	**72.07±0.84**
	1-OE	26.14±3.41	31.28±3.19	39.21±0.21	31.84±4.21	35.22±4.83	37.47±1.53	26.53±0.71	**39.35±1.83**	37.16±1.90
	1-Cov	58.13±0.78	58.76±1.20	59.06±0.93	56.94±1.52	56.55±0.89	59.19±0.81	53.99±0.39	58.02±1.97	**65.64±1.21**
	AP	34.45±1.38	37.52±1.14	40.18±0.65	36.57±2.03	37.16±2.17	39.83±0.18	27.49±0.35	41.41±0.68	**43.86±1.15**
	AUC	67.92±0.80	68.84±1.00	68.94±0.87	66.52±1.41	65.81±0.49	69.76±0.68	62.29±0.39	68.55±2.08	**74.65±1.01**
	Ave.R	6.83	5.17	2.50	6.50	6.50	3.67	8.67	3.50	**1.67**
Espgame	1-HL	97.88±0.12	97.50±0.56	98.40±0.00	98.26±0.01	**98.41±0.00**	79.07±0.04	94.75±4.96	97.78±0.18	98.25±0.01
	1-RL	62.26±0.32	63.18±0.50	63.94±0.30	62.14±0.84	64.67±0.66	62.57±0.98	54.05±0.64	66.66±0.60	**69.79±0.47**
	1-OE	14.49±1.20	15.75±1.08	14.97±0.70	13.63±0.91	18.75±0.24	10.63±0.54	17.27±7.21	1.48±1.19	**31.77±1.51**
	1-Cov	29.73±0.31	30.41±0.49	33.06±0.30	31.95±0.66	34.65±0.60	30.14±1.01	25.73±0.17	33.37±0.43	**35.80±0.55**
	AP	9.83±0.35	10.78±0.32	11.28±0.13	10.19±0.47	13.38±0.29	9.66±0.55	5.82±0.09	6.77±0.78	**19.13±0.58**
	AUC	62.15±0.32	63.35±0.56	64.12±0.23	62.20±0.90	66.28±0.70	62.65±0.91	56.03±0.06	66.64±0.79	**69.86±0.54**
	Ave.R	6.67	5.17	3.67	5.83	2.17	7.17	7.83	5.00	**1.50**
IAPR TC-12	1-HL	97.54±0.04	97.67±0.18	98.00±0.00	97.93±0.03	98.00±0.00	79.41±0.02	88.22±1.24	97.18±0.14	**98.04±0.02**
	1-RL	61.86±0.32	65.48±2.45	66.90±0.14	65.17±0.91	69.20±0.33	70.78±0.57	55.79±0.91	69.68±0.73	**72.66±0.48**
	1-OE	14.86±0.64	18.27±1.68	29.51±0.81	27.42±5.44	33.89±4.67	23.45±4.16	15.02±9.56	1.49±1.39	**34.33±1.37**
	1-Cov	25.98±0.43	28.56±1.95	29.24±0.31	28.12±1.16	31.06±0.50	34.23±1.08	23.23±0.29	33.67±0.46	**35.40±0.53**
	AP	9.18±0.24	12.07±1.57	15.10±0.10	13.42±0.78	17.37±0.89	14.88±0.63	6.25±0.07	7.32±0.43	**19.66±0.07**
	AUC	62.12±0.24	65.52±2.38	66.95±0.09	65.38±0.67	69.48±0.31	70.87±0.60	57.53±0.11	69.75±0.88	**72.56±0.50**
	Ave.R	7.50	5.83	3.83	5.67	3.17	4.00	8.50	5.50	**1.00**
MIR FLICKR	1-HL	86.63±0.25	85.90±0.87	87.46±0.07	86.24±0.31	87.12±0.19	73.17±0.10	79.90±0.13	86.24±0.94	**87.61±0.09**
	1-RL	75.11±0.43	76.91±1.61	76.23±0.20	73.11±0.50	74.97±0.95	75.97±0.88	60.07±0.19	79.74±1.68	**82.34±0.75**
	1-OE	44.98±0.76	43.88±1.76	36.58±0.73	35.51±1.19	35.78±1.07	36.97±2.96	42.44±0.37	38.67±6.72	**61.34±0.99**
	1-Cov	46.06±0.59	48.98±2.79	49.39±0.23	46.30±0.49	47.18±3.13	48.15±2.14	35.08±0.09	53.12±4.12	**56.94±0.98**
	AP	42.55±0.42	44.11±0.79	40.80±0.20	38.50±0.18	40.71±0.22	42.18±0.56	29.72±0.34	45.69±2.73	**54.89±0.94**
	AUC	73.96±0.41	75.67±1.91	75.42±0.17	72.27±0.58	73.97±1.38	74.62±0.72	62.80±0.14	79.16±1.28	**81.14±0.93**
	Ave.R	5.17	3.83	4.33	7.50	6.17	5.83	8.00	3.17	**1.00**

ing them with average imputation, followed by feature concatenation. All parameters of compared methods are configured as the recommended values in their original codes.

Implementation Details. Each dataset is divided into training, validation and test sets in the ratio of 7:1:2. To simulate the partial view setting, a certain proportion of instances defined by the partial example ratio (PER) are randomly selected as unavailable instances on each view. Besides, we ensure each sample contains at least one complete view to avoid invalid cases. In the multi-label noise configuration, we set a uniform noise rate for each label, i.e, ρ of positive and negative labels are flipped to their opposite. All results are obtained from ten independent runs, with final outcomes reported as mean values alongside standard deviations. Our model is implemented by PyTorch on one NVIDIA GeForce RTX 4090 GPU of 24GB memory.

3.2. Experimental Results.

Performance Evaluation. To evaluate the capabilities of DMMIvNL in addressing data deficiencies, we conduct extensive comparative experiments focusing on two critical aspects, i.e, label noise and view missing. Besides, we assign a broad range of noise rates and missing proportions to demonstrate the versatility of our approach. For label noise, ρ is considered at levels of $\{10\%, 15\%, 20\%, 25\%, 30\%, 35\%, 40\%\}$,

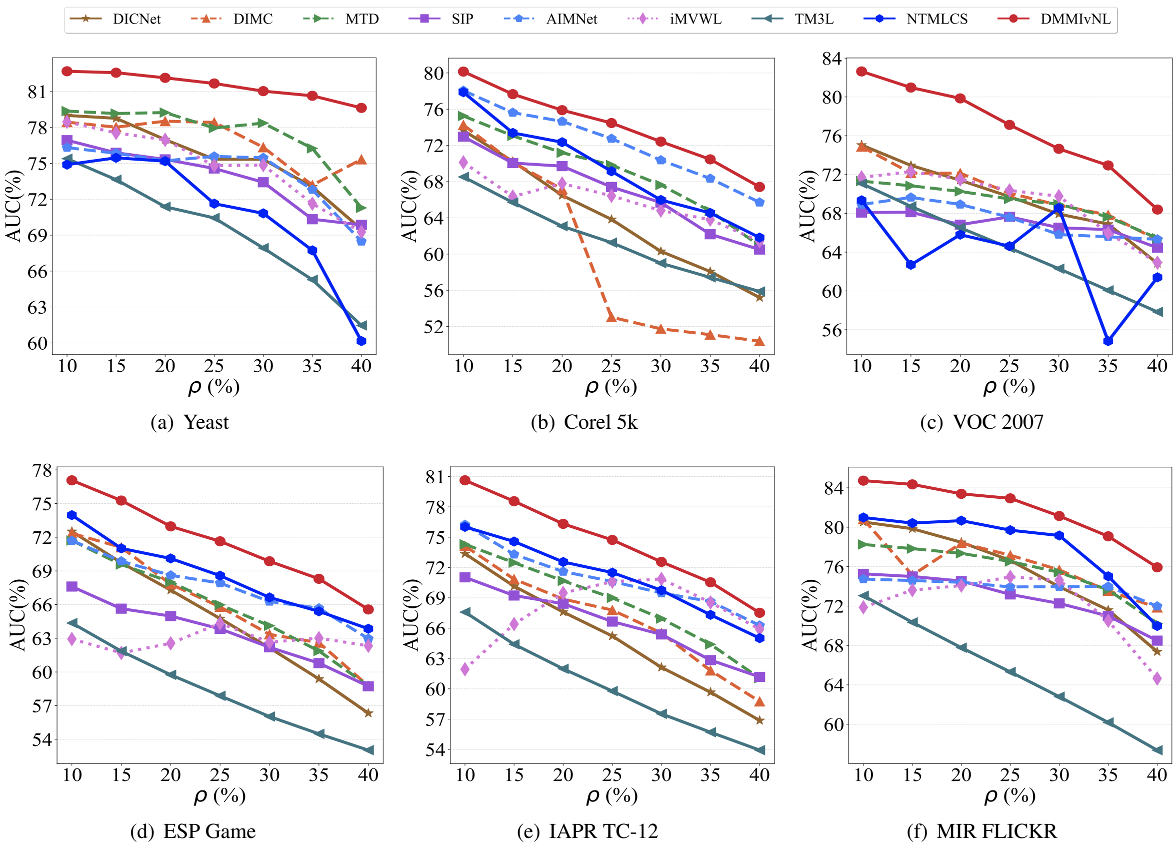


Figure 2. AUC comparisons on six datasets with noise rate ρ varying from 10% to 40% while PER=50%.

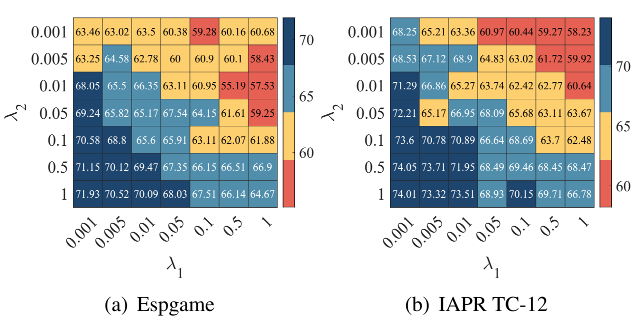
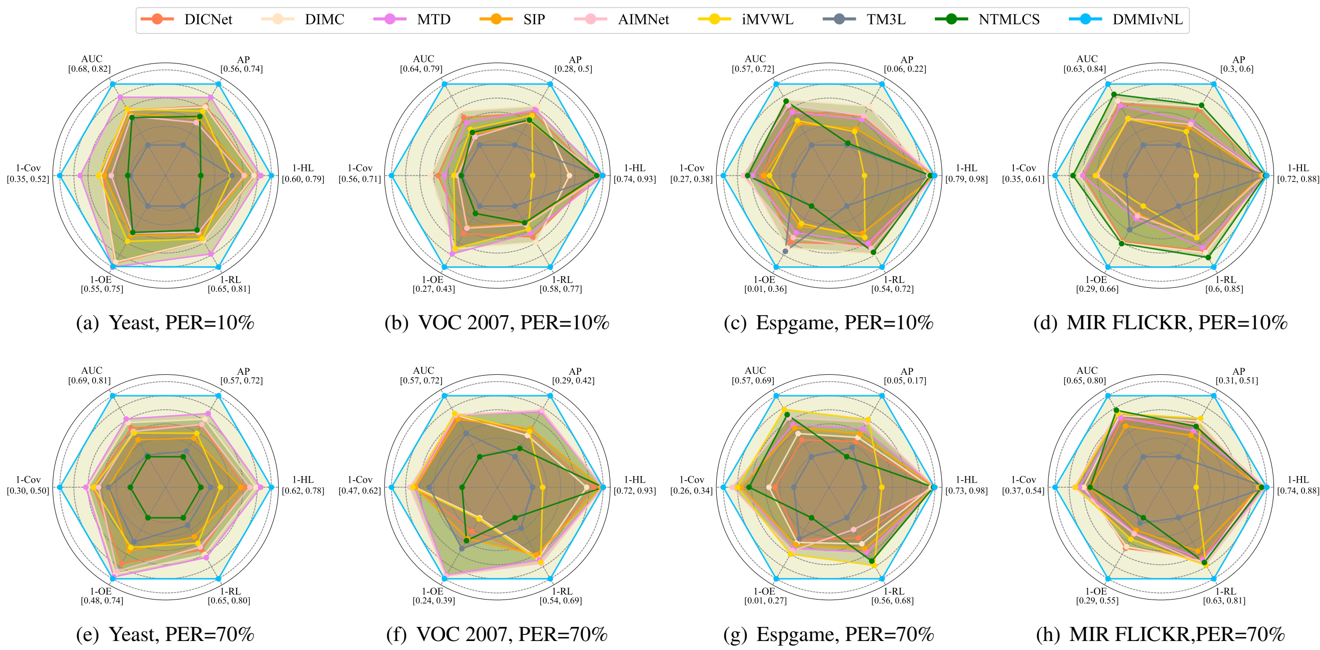


Figure 3. Parameter analysis of the trade-off parameters λ_1 and λ_2 on Espgame and IAPR TC-12.

while for view missing, PER encompasses values of $\{10\%, 30\%, 50\%, 70\%\}$. Tab. 1 displays all metrics with ρ fixed at 30% and PER fixed at 50%, Fig. 2 illustrates the variation in AUC as ρ varies from 10% to 40%, and Fig. 4 shows the metrics distribution when PER=10% and 70%. Additional results are included in the Appendix.

Regarding the label noise, we can find that 1) Our method outperforms all compared approaches, particularly distinguished in the most representative AP metric. For instance, our method achieves a remarkable AP improvement of up to 43% on Espgame in Tab. 1. 2) Compared to NTMLCS for handling label noise, our approach reveals substantial advantages. Besides, in terms of the performance on VOC 2007, DMMIvNL exhibits greater stability than the stepwise estimation way of NTMLCS, which is achieved through the cycle-consistency framework that simultaneously determines the transition matrix and classifier. Additionally, our approach surpasses methods developed for incomplete MvMLC. For example, Fig. 2 illustrates that our DMMIvNL reliably improves performance even against the top-performing MvMLC methods MTD and AIMNet. 3) As ρ increases from 10% to 40%, DMMIvNL consistently outperforms the other eight methods. Moreover, many methods lose reliability at elevated noise levels, such as NTMLCS on Yeast, TM3L on MIR FLICKR and DIMC on Corel 5k, while our approach continues to learn a competent classifier and deliver robust performance.

As for the view missing, we have the following insights: 1) Our method exhibits outstanding performance in handling missing views. As illustrated in the radar chart of Fig. 4, DMMIvNL consistently occupies the outermost circle, reflecting impressive results across all metrics. 2)



(a) Yeast, PER=10% (b) VOC 2007, PER=10% (c) Espgame, PER=10% (d) MIR FLICKR, PER=10%

(e) Yeast, PER=70% (f) VOC 2007, PER=70% (g) Espgame, PER=70% (h) MIR FLICKR,PER=70%

Figure 4. Experimental results of nine methods on four datasets with PER=10% and 70% while $\rho = 30\%$.

As PER changes from 10% to 70%, DMMIvNL continues to achieve top performance. Notably, while DIMC experiences a substantial decline at 70% PER, our method remains efficient. It indicates that DMMIvNL delivers strong robustness across varying levels of feature absence. 3) Our method makes remarkable breakthroughs over deep learning based methods for missing views, like SIP, MTD, and DICNet. For example, on MIR FLICKR, DMMIvNL improves 1-OE by nearly 50% compared to SIP at 50% PER. This improvement is primarily due to the focus on extracting the shared semantics and specific features through information separation, an aspect overlooked by other methods.

Ablation Study and Parameter Sensitivity. The ablation experiments are conducted to analyze the effect of the three crucial aspects of DMMIvNL, i.e., representation extraction based on the information bottleneck theory ($\mathcal{L}_{IB}$), cycle-consistency estimation for noise identification and classifier learning ($\mathcal{L}_{ID}$), and model regularization to combat overfitting ($\mathcal{L}_r$). Note that loss $\mathcal{L}_c$ is retained since it is essential for classifier improvement during early training. The first three rows of Tab. 2 indicate that $\mathcal{L}_{IB}$, $\mathcal{L}_{ID}$ and $\mathcal{L}_r$ are integral components of DMMIvNL. Besides, the comparison between the empirical loss $\mathcal{L}_{vol}$ from problem (11) and $\mathcal{L}_{ID}$ confirms that the cycle-consistency framework enhances the stability of transition matrix estimation. To study the impact of λ_1 and λ_2, we present the results under distinct parameter settings in Fig. 3. As shown in the heatmap, the performance of DMMIvNL exhibits variability and improves as λ_1 gets closer to 0.001 and λ_2 approaches 1.

Table 2. Ablation study on Yeast, Corel 5k and VOC 2007 with PER=50% and $\rho = 30\%$. '✓' and '✗' represent the used and not used corresponding item, respectively.

$\mathcal{L}_{vol}$	$\mathcal{L}_{IB}$	$\mathcal{L}_{ID}$	$\mathcal{L}_r$	Espgame		Corel 5k		VOC 2007	
				AP	AUC	AP	AUC	AP	AUC
✗	✗	✓	✓	0.1750	0.6846	0.1713	0.6972	0.4145	0.7366
✗	✓	✗	✓	0.1838	0.7113	0.1850	0.7089	0.4305	0.7367
✗	✓	✓	✗	0.1834	0.7131	0.1800	0.7096	0.4359	0.7148
✓	✗	✓	✓	0.1499	0.6732	0.1543	0.6284	0.3612	0.6715
✗	✓	✓	✓	**0.1934**	**0.7307**	**0.1997**	**0.7284**	**0.4622**	**0.7566**

4. Conclusion

To address the dual challenges of view missing and multi-label noise, we propose DMMIvNL in this paper. Built upon the information bottleneck theory, a mutual information model is established to constrain information exchange across diverse representations, along with a derived theoretical upper bound. Besides, we prove that minimizing the volume of the transition matrix for each category consistently leads to the optimal classifier. Moreover, stability in noise recognition is further improved by introducing a cycle-consistency estimation framework, while raw fidelity information and inherent label correlations are used to reduce overfitting. Finally, extensive experiments demonstrate the superiority of DMMIvNL. In future work, we will leverage the prior knowledge embedded in LLM to rectify inaccurate labels and tackle the issue of class imbalance.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China [Grant No. 62376281], and the Key NSF of China under Grant No. 62136005.

References

[1] Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *SIGCHI Conference on Human Factors in Computing Systems*, page 319–326, 2004. 5

[2] Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence scores make instance-dependent label-noise learning possible. In *ICML*, pages 825–836, 2021. 4

[3] Jia-Yao Chen, Shao-Yuan Li, Sheng-Jun Huang, and Chm. Unm: A universal approach for noisen, songcan and wang, lei and xie, ming-kun. *IEEE Transactions on Knowledge and Data Engineering*, 36:4968–4980, 2024. 2

[4] De Cheng, Yixiong Ning, Nannan Wang, Xinbo Gao, Heng Yang, Yuxuan Du, Bo Han, and Tongliang Liu. Class-dependent label-noise learning with cycle-consistency regularization. pages 11104–11116, 2022. 4

[5] Pinar Duygulu, Kobus Barnard, Joao FG de Freitas, and David A Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, pages 97–112, 2002. 5

[6] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010. 5

[7] Yuling Fan, Jinghua Liu, Jianeng Tang, Peizhong Liu, Yaojin Lin, and Yongzhao Du. Learning correlation information for multi-label feature selection. *Pattern Recognition*, 145: 109899, 2024. 4

[8] Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *ICLR*, 2020. 2

[9] Xiao Fu, Kejun Huang, and Nicholas D Sidiropoulos. On identifiability of nonnegative matrix factorization. *IEEE Signal Processing Letters*, 25(3):328–332, 2018. 4

[10] Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop onto Image*, pages 1–11, 2006. 5

[11] Cordelia Guillaumin. Multimodal semi-supervised learning for image classification. In *CVPR*, pages 902–909, 2010. 5

[12] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, I Tsang, and Masashi Sugiyama. Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, page 4, 2020. 4

[13] Pingting Hao, Weiping Ding, Wanfu Gao, and Jialong He. Exploring view-specific label relationships for multi-view multi-label feature selection. *Information Sciences*, 681: 121215, 2024. 1

[14] Huaxi Huang, Hui Kang, Sheng Liu, Olivier Salvado, Thierry Rakotoarivelo, Dadong Wang, and Tongliang Liu. Paddles: Phase-amplitude spectrum disentangled early stopping for learning with noisy labels. In *ICCV*, pages 16719–16730, 2023. 2

[15] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *ICMR*, pages 39–43, 2008. 5

[16] Xiaodong Jia, Xiao-Yuan Jing, Xiaoke Zhu, Songcan Chen, Bo Du, Ziyun Cai, Zhenyu He, and Dong Yue. Semi-supervised multi-view deep discriminant representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2496–2509, 2020. 1, 4

[17] Zhangqi Jiang, Tingjin Luo, and Xinyan Liang. Deep incomplete multi-view learning network with insufficient label information. In *AAAI*, pages 12919–12927, 2024. 2

[18] Rihuan Ke. Deep variation prior: Joint image denoising and noise variance estimation without clean data. *IEEE Transactions on Image Processing*, 33:2908–2923, 2024. 4

[19] Shikun Li, Xiaobo Xia, Hansong Zhang, Yibing Zhan, Shiming Ge, and Tongliang Liu. Estimating noise transition matrix with label correlations for noisy multi-label learning. In *NeurIPS*, pages 24184–24198, 2022. 2

[20] Xuefeng Li, Tongliang Liu, Bo Han, Gang Niu, and Masashi Sugiyama. Provably end-to-end label-noise learning without anchor points. In *ICML*, pages 6403–6413, 2021. 4

[21] Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4447–4461, 2022. 3

[22] Chengliang Liu, Jie Wen, Xiaoling Luo, Chao Huang, Zhihao Wu, and Yong Xu. Dicnet: Deep instance-level contrastive network for double incomplete multi-view multi-label classification. In *AAAI*, pages 8807–8815, 2023. 1

[23] Chengliang Liu, Jie Wen, Xiaoling Luo, and Yong Xu. Incomplete multi-view multi-label learning via label-guided masked view-and category-aware transformers. In *AAAI*, pages 8816–8824, 2023. 1, 5

[24] Chengliang Liu, Jinlong Jia, Jie Wen, Yabo Liu, Xiaoling Luo, Chao Huang, and Yong Xu. Attention-induced embedding imputation for incomplete multi-view partial multi-label classification. In *AAAI*, pages 13864–13872, 2024. 1

[25] Chengliang Liu, Jie Wen, Yabo Liu, Chao Huang, Zhihao Wu, Xiaoling Luo, and Yong Xu. Masked two-channel decoupling framework for incomplete multi-view weak multi-label learning. In *NeurIPS*, 2024. 1, 4, 5

[26] Chengliang Liu, Gehui Xu, Jie Wen, Yabo Liu, Chao Huang, and Yong Xu. Partial multi-view multi-label classification via semantic invariance learning and prototype modeling. In *ICML*, 2024. 1, 2

[27] Chengliang Liu, Jie Wen, Yong Xu, Bob Zhang, Liqiang Nie, and Min Zhang. Reliable representation learning for incomplete multi-view missing multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–17, 2025. 1

[28] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, 2015. 4

[29] Tingjin Luo, Yang Yang, Dongyun Yi, and Jieping Ye. Robust discriminative feature learning with calibrated data reconstruction and sparse low-rank model. *Applied Intelligence*, 54(3):2867–2880, 2024. 1

[30] Yong Luo, Dacheng Tao, Chang Xu, Chao Xu, Hong Liu, and Yonggang Wen. Multiview vector-valued manifold regularization for multilabel image classification. *IEEE Transactions on Neural Networks and Learning Systems*, 24(5): 709–722, 2013. 1

[31] Yubin Park and Joyce C Ho. Tackling overfitting in boosting for noisy healthcare data. *IEEE Transactions on Knowledge and Data Engineering*, 33(7):2995–3006, 2019. 1

[32] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, pages 1944–1952, 2017. 4

[33] Yalan Qin, Hanzhou Wu, Xinpeng Zhang, and Guorui Feng. Semi-supervised structured subspace learning for multi-view clustering. *IEEE Transactions on Image Processing*, 31:1–14, 2021. 1

[34] Alisa Smirnova, Jie Yang, Dingqi Yang, and Philippe Cudre-Mauroux. Nessy: A neuro-symbolic system for label noise reduction. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8300–8311, 2022. 1

[35] Shiding Sun, Bo Wang, and Yingjie Tian. Decoupled representation for multi-view learning. *Pattern Recognition*, 151: 110377, 2024. 2

[36] Yu-Yin Sun, Yin Zhang, and Zhi-Hua Zhou. Multi-label learning with weak label. In *AAAI*, pages 593–598, 2010. 1

[37] Qiaoyu Tan, Guoxian Yu, Carlotta Domeniconi, Jun Wang, and Zili Zhang. Incomplete multi-view weak-label learning. In *IJCAI*, pages 2703–2709, 2018. 1

[38] Ziyi Wang and Yitian Xu. A two-stage multi-view partial multi-label learning for enhanced disambiguation. *Knowledge-Based Systems*, 293:111680, 2024. 1

[39] Jie Wen, Chengliang Liu, Shijie Deng, Yicheng Liu, Lunke Fei, Ke Yan, and Yong Xu. Deep double incomplete multi-view multi-label learning with incomplete labels and missing views. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):11396–11408, 2023. 1

[40] Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning with noisy label identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3676–3687, 2021. 1

[41] Wulin Xie, Xiaohuan Lu, Yadong Liu, Jiang Long, Bob Zhang, Shuping Zhao, and Jie Wen. Uncertainty-aware pseudo-labeling and dual graph driven network for incomplete multi-view multi-label classification. In *ACMMM*, pages 6656–6665, 2024. 1

[42] Yu Yao, Tongliang Liu, Mingming Gong, Bo Han, Gang Niu, and Kun Zhang. Instance-dependent label-noise learning under a structural causal model. In *NeurIPS*, 2021. 2

[43] Yitong Yao, Jing Zhang, Peng Zhang, and Yueheng Sun. A dual-branch learning model with gradient-balanced loss for long-tailed multi-label text classification. *ACM Transactions on Information Systems*, 42(2):1–24, 2023. 5

[44] Selim F Yilmaz, E Batuhan Kaynak, Aykut Koç, Hamdi Dibeklioğlu, and Suleyman Serdar Kozat. Multi-label sentiment analysis on 100 languages with dynamic weighting for label imbalance. *IEEE Transactions on Neural Networks and Learning Systems*, 34:331–343, 2021. 4

[45] Dawei Zhao, Qingwei Gao, Yixiang Lu, and Dong Sun. Two-step multi-view and multi-label learning with missing label via subspace learning. *Applied Soft Computing*, 102:107120, 2021. 5

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

[47] Wenzhang Zhuge, Tingjin Luo, Ruidong Fan, Hong Tao, Chenping Hou, and Dongyun Yi. Absent multiview semisupervised classification. *IEEE Transactions on Cybernetics*, 54(3):1708–1721, 2024. 1