Double-Layer Hybrid-Label Identification Feature Selection for Multi-View Multi-Label Learning

Pingting Hao[1,2], **Kunpeng Liu**[3], **Wanfu Gao**[1,2*]

[1] College of Computer Science and Technology, Jilin University, China
[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China
[3] Department of Computer Science, Portland State University, Portland, OR 97201 USA
haopingting@jlu.edu.cn, kunpeng@pdx.edu, gaowf@jlu.edu.cn

Abstract

Multi-view multi-label feature selection aims to select informative features where the data are collected from multiple sources with multiple interdependent class labels. For fully exploiting multi-view information, most prior works mainly focus on the common part in the ideal circumstance. However, the inconsistent part hidden in each view, including noises and specific elements, may affect the quality of mapping between labels and feature representations. Meanwhile, ignoring the specific part might lead to a suboptimal result, as each label is supposed to possess specific characteristics of its own. To deal with the double problems in multi-view multi-label feature selection, we propose a unified loss function which is a totally splitting structure for observed labels as hybrid labels that is, common labels, view-to-all specific labels and noisy labels, and the view-to-all specific labels further splits into several specific labels of each view. The proposed method simultaneously considers the consistency and complementarity of different views. Through exploring the feature weights of hybrid labels, the mapping relationships between labels and features can be established sequentially based on their attributes. Additionally, the interrelatedness among hybrid labels is also investigated and injected into the function. Specific to the specific labels of each view, we construct the novel regularization paradigm incorporating logic operations. Finally, the convergence of the result is proved after applying the multiplicative update rules. Experiments on six datasets demonstrate the effectiveness and superiority of our method compared with the state-of-the-art methods.

Introduction

Multi-view multi-label learning has attracted much attention with the rapid growth of the data source (Fu et al. 2022; Liu et al. 2023b). It is an extension of multi-label learning from a single view to many views (at least two views). From the perspective of information completeness, multi-view multi-label learning possesses advantages due to its rich semantic content. However, the explosive data brings more noisy, irrelevant and redundant features (Luo et al. 2017; Chang and Yang 2016). Thus, the task about how to explore the informative features is still urgent in the field of multi-view multi-label learning (MVML), namely multi-view multi-label feature selection, which makes the model cost-effective and provides accurate data representation (Lin et al. 2022; Cui et al. 2021).

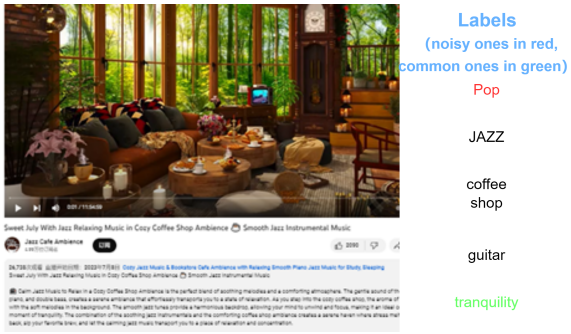


Figure 1: An illustration to describe the multi-view multi-label problem. The webpage can be represented from different views such as video, text and audio. The red one represents the noisy label and the green one is owned by all views.

Currently, the basic assumption in traditional MVML methods is that the relevant label of each instance has been annotated precisely (Zhu, Li, and Zhang 2015). Rather than the ideal circumstance, the noises actually exist among annotations for labels. Meanwhile, the fact is that the label of each view is not completely identical in real-world scenarios. As an illustration, a webpage may contain three distinct types of views, comprising video, text, and audio, together with several labels provided by annotators. For instance, the music style can be roughly labeled as "Pop", which should in fact be labeled as "JAZZ". While "tranquility" is a shared label among all three types of views, "guitar" is exclusively relevant to the audio view and is not applicable to the video or text views.

Distinct from unsupervised learning, accurately incorporating labeled data into the learning process remains a fundamental challenge in the context of multi-view multi-learning problem. Existing methods aim to refrain from imposing the basic assumption while simultaneously addressing the issue of the label noise. The consistency part is effectively acquired through dimension reduction to the label matrix (Zhang et al. 2020b), or considers the relation between labels to filter the noises (Liu et al. 2023a). Consequently, the

*correponding author

complementary information across views is not given much consideration, and the extension of the method to encompass non-aligned views is restricted. Thus, the view-specific label is defined and denotes explicitly that the insufficient consideration of the relationship between the features and labels of each view leads to the unideal learning effect (Zhu et al. 2020; Zhao et al. 2022a). Regrettably, due to the existing relationship between view-specific labels and observed labels, the dimension reduction for view-specific labels is not enough for the direction of filtering noises. The feature weight is further influenced by the 'FP' or 'FN' relationship between labels and features. The 'FP' relationship means that the noises still exist, and 'FN' means that the ground-truth label is filtered by mistake. Thus, the derivation disturbs the order of features. How to identify noises in the label matrix from the scenario with more dimensions which has complex relationships is still a challenging problem.

Therefore, this paper attempts to approach the issue of label noise in multi-view multi-label feature selection from a different perspective, and then to identify the common and specific labels of each view simultaneously. Drawing on the analysis presented above, it is evident that the difficulty in identifying noise lies in the fact that it lacks a precisely defined interpretation. Therefore, the resolution approach commonly employed is to utilize the principle of "more is better than less", whereby the minority is treated as noise, i.e., the low-rank technique (Yu et al. 2018; Jian et al. 2016), the filtering rule based on the order of the correlation level (Gong, Yuan, and Bao 2021; González-López, Ventura, and Cano 2019). In fact, the noises can be generated by establishing a constant relationship with the observed variables from another perspective. On this basis, we propose a method named *Double-layer Hybrid-Label Identification* (DHLI) regardless of noise pattern to improve the feature selection procedure in the multi-view multi-label scenario. According to the various object-oriented approaches, the type of labels is divided into two layers, while establishing connections among intra-layer and inter-layer separately. Compared with existing methods, the method DHLI we propose in this paper has the following main contributions:

- The method DHLI provides a novel splitting structure of the observed label, which has the potential to avoid the noises and preserve the common and complementary information;
- We devise a novel regularization paradigm incorporating logic operations to constrain the learning direction of variables in the optimization process, and the multiplicative update rules act on the optimization process with proven convergence;
- We conduct extensive experiments on six datasets, and the results demonstrate the efficiency and excellent performance of the proposed method.

Related Work

Multi-Label Feature Selection

Different from the single label, one instance is assigned a set of proper labels to explicitly express multiple semantic meanings (Zhang et al. 2020a). Thus, it is essential to facilitate the learning process by exploiting correlations among labels (Zhang and Zhou 2013). According to the degree of label correlation exploitation (Zhang and Zhang 2010), the multi-label feature selection can be divided into three strategies. The first-order strategy (Lee and Kim 2015, 2017; Lin et al. 2015) uses the accumulated mutual information between candidate features and each label, without considering the label correlation. The second-order strategy focuses on the relationship between pair of labels. One of the related techniques usually adopts information theory to measure the dependency between labels and features (Zhang, Liu, and Gao 2019; Lim and Kim 2020) or to compute the label weight for each label prior to linear weighted sum method (Xiong et al. 2021; Qian et al. 2022). Other techniques such as manifold learning (Zhang et al. 2020a) are to preserve the similarity of the structure between two matrices. Similarly, the label correlation could also be calculated by heat kernel (Zhang et al. 2019) or hamming distance (Lin et al. 2021). Unfortunately, the complex relationships among labels often appear in the real world (Eswaran, Kumar, and Faloutsos 2020).

Thus, to excavate more effective information and remove unnecessary dependency relationships, the high-order strategy is taken into the relationship among labels. Sparse learning (Jian et al. 2016; Huang et al. 2016; Fan et al. 2021) and causal mechanism (Wu et al. 2020; Yu et al. 2021) are the prevalent approaches in high-order strategies. Along with it comes the consideration of data relations as dimensionality increases from single view to multi-view, the methods employed for multi-label data are not directly applicable to multi-view scenarios.

Multi-View Multi-Label Learning

Multi-view multi-label learning has drawn extensive research enthusiasm in recent years. Some methods concentrate on learning the shared components among all views (Liu et al. 2015; Luo et al. 2013; Yin and Zhang 2023). A subsequent work design variables of different views to explore more complementary information (Zhang et al. 2018; Wu et al. 2019; Lyu et al. 2022; Tan et al. 2019). Actually, the challenges for MVML inevitably occur in the cases such as missing labels, incomplete views and non-aligned views. However, the third case is always ignored or hard to deal with (Li and Chen 2021). To overcome this challenge, some recent works (Zhao et al. 2022a,b) find the importance of the information for view specific and propose to learn a new variable named view-specific labels, which means each view has an individual label matrix including the common part.

Despite the inclusion of finer-grained hierarchical label structures in these methods, there has been a lack of research examining the interrelations between labels. This is particularly relevant given the noise challenges that can arise both across and within different hierarchical levels. To fill the gap, this paper attempts to explore more complex relationships to build a robust mapping between label and feature representation in multi-view multi-label learning.

The Proposed Method

Problem Settings

Unless otherwise stated, the main notations used in this paper are listed as follows. Suppose $X = \left\{X^{(i)}\right\}_{i=1}^{V}$ is d-dimensional feature vector and contains V different views, and $X^{(i)} \in \mathbb{R}^{n\times d(i)}$ where $x_{(m)}^{(i)} \in X^{(i)}$ is a $d(i)$-dimensional feature vector and denotes the m-th instance for i-th view. Also, the Frobenius norm is denoted as $\left\|X^{(i)}\right\|_F = \sqrt{\sum_{p=1}^{n}\sum_{q=1}^{d(i)}\left(x_{(pq)}^{(i)}\right)^2}$, and the $l_{2,1}$-norm is denoted as $\left\|X^{(i)}\right\|_{2,1} = \sum_{p=1}^{n}\sqrt{\sum_{q=1}^{d(i)}\left(x_{(pq)}^{(i)}\right)^2}$. $Y \in \{0,1\}^{n\times c}$ represents the label assignments for the corresponding labeled instances, and the j-th label is denoted by $y_{(j)}$. If $x_{(m)}^{(i)}$ is associated with the j-th label, then $y_{(mj)}$ is set 1; otherwise, $y_{(mj)}$ is set 0.

Definition 1 *(First-layer label correlation) Given a multi-view multi-label data set, $Y_c \in \{0,1\}^{n\times c}$, $Y_s \in \{0,1\}^{n\times c}$ and $Y_n \in \{0,1\}^{n\times c}$ denote common labels, specific labels and noisy labels for all views. The relationship between the matrices above and the observed label matrix is defined as follows:*

$$Y = Y_c \otimes Y_s \otimes Y_n, \tag{1}$$

where the OR operation between matrices is denoted by '$\otimes$', and the computation is performed on an element-by-element basis.

Definition 2 *(Second-layer label correlation) The specific part of each view for label matrix is defined as $\left\{y_s^{(i)}\right\}_{i=1}^{V} \in \{0,1\}^{n\times c}$ where V is the number of views.*

$$Y_s = [y_s^{(1)}, y_s^{(2)}, ..., y_s^{(V)}]. \tag{2}$$

Problem Formulation

Following the traditional multi-label feature selection model, the minimum regression error across the data can be conducted as:

$$\min_{W} \|XW - Y\|_F^2 + \Phi(W), \tag{3}$$

where $W \in \mathbb{R}^{d\times c}$ denotes the feature weight, and the $\Phi(W)$ denotes the regularization paradigm of W to control the model complexity.

The expansion of Formula (3) into the multi-view multi-label scenario, the feature weight of each view can be required by:

$$\min_{W^{(i)}} \sum_{i=1}^{V}\left\|X^{(i)}W^{(i)} - Y\right\|_F^2 + \Phi(W), \tag{4}$$

where $W^{(i)} \in \mathbb{R}^{d(i)\times c}$ denotes the feature weight of each view, and the second term continues to aim at all the features. Achieving accurate learning of feature weights is dependent on the variable for the label matrix in the first term. This variable plays a crucial role in shaping the quality of the mapping between label spaces and feature representations.

As aforementioned, Definition 1 offers an alternative perspective where the inherent noise in the observed label matrix can be filtered out. Meanwhile, this also preserves the distinct attributes of each view. This entails the division of feature weights into two distinct types, namely the common feature weight and the specific feature weight for each view, and Formula (4) is rewritten for different types of mapping as:

$$\begin{aligned}\min_{W^{(i)},U^{(i)},Y_c,Y_s} &\sum_{i=1}^{V}\left\|X^{(i)}W^{(i)} - Y_c\right\|_F^2 \\ &+ \sum_{i=1}^{V}\left\|X^{(i)}U^{(i)} - Y_s\right\|_F^2 + \Phi(W) + \Omega(U),\end{aligned} \tag{5}$$

where $W^{(i)} \in \mathbb{R}^{d(i)\times c}$ denotes the common feature weight of each view, and $U^{(i)} \in \mathbb{R}^{d(i)\times c}$ is the specific feature weight for i-th view. The function for the sparsity of the specific feature weight matrix is defined as $\Omega(U)$.

Distinct from the characteristics of the "common", the element of "specific" is primarily focused on each individual view. It suggests that Y_s is not uniform across views and instead varies from one view to another. Hence, a specific label for each view is explicated through Definition 2, and we utilize the $l_{2,1}$-norm regularization to generate the group sparsity (Yuan and Lin 2006). Therefore, the Formula (6) can be written as:

$$\begin{aligned}\min_{W^{(i)},U^{(i)},Y_c,y_s^{(i)}} &\sum_{i=1}^{V}\left\|X^{(i)}W^{(i)} - Y_c\right\|_F^2 \\ &+ \sum_{i=1}^{V}\left\|X^{(i)}U^{(i)} - y_s^{(i)}\right\|_F^2 + \|W\|_{2,1} + \|U\|_{2,1}.\end{aligned} \tag{6}$$

Leveraging multi-view data can facilitate a more comprehensive and precise representation of objects. This underscores the significance of identifying the commonality between views. Thus, we assume that the common part accounts for the majority of the observed label matrix. Then, we define '$\ominus$' as a metric to gauge the dissimilarity between two values appearing in identical locations across matrices. This metric is formally defined as follows:

$$y_{(pq)} \ominus y_{c(pq)} = \begin{cases} 1 & y_{(pq)} \neq y_{c(pq)} \\ 0 & y_{(pq)} = y_{c(pq)} \end{cases}. \tag{7}$$

Furthermore, it is worth noting that this symbol can also be extended to accommodate operations across matrices. With the assumption of sparse label noise, the l_1-norm regularization technique is used to enforce element-wise sparsity (Efron et al. 2004; Zhu et al. 2013). Consequently, this results in the formulation of the related Formula as follows:

$$\min_{Y_c,Y_n} \|Y \ominus Y_c\|_F^2 + \|Y_n\|_1. \tag{8}$$

The definition of $y_s^{(i)}$ pertains to situations where all views lack a shared label. This represents a significant deviation from the common component. In an effort to mitigate this challenge, we introduce a logic-based imposition whereby elements located in the same position of different views are constrained.

$$\sum_{p=1}^{n}\sum_{q=1}^{c} y_{s(pq)}^{(1)} \oplus y_{s(pq)}^{(2)} \cdots \oplus y_{s(pq)}^{(V)} = 0, \tag{9}$$

where $y_{s(pq)}^{(i)}$ denotes the specific value of q-th label for p-th instance in i-th view. The operation described differs from the definition of Y_c. It is defined with '$\oplus$' such that if all elements are 1, the result is 1; otherwise, the result is 0. As a consequence, the specific label matrix for each view can be restricted in the following manner:

$$\min_{y_s^{(i)}} \left\| y_s^{(1)} \oplus y_s^{(2)} ... \oplus y_s^{(V)} \right\|_F^2. \tag{10}$$

Another characteristic of $y_s^{(i)}$ is the existence of a relationship between $y_s^{(i)} (i = 1, 2, \ldots, V)$ and Y_s, as stipulated in Definition 2. Specifically, it can be expressed that if a label of an instance is marked within $y_s^{(i)}$, then the label of that instance should be designated in the same fashion in Y_s, which in turn constrains the position where $y_s^{(i)}$ is marked, in order to resist the $y_s^{(i)}$ to be zero matrix when Y_s is not a zero matrix. Drawing upon this description, the Formula can be expressed as follows:

$$\min_{y_s^{(i)}, Y_s} \left\| y_s^{(1)} \otimes y_s^{(2)} ... \otimes y_s^{(V)} \ominus Y_s \right\|_F^2. \tag{11}$$

Thus, the overall loss of our method DHLI can be calculated as:

$$\begin{aligned}
&\min_{W^{(i)}, U^{(i)}, Y_c, Y_s, Y_n, y_s^{(i)}} \sum_{i=1}^{V} \left\| X^{(i)} W^{(i)} - Y_c \right\|_F^2 \\
&+ \sum_{i=1}^{V} \left\| X^{(i)} U^{(i)} - y_s^{(i)} \right\|_F^2 + \alpha \left\| Y \ominus Y_c \right\|_F^2 + \beta \left\| Y_n \right\|_1 \\
&+ \gamma (\left\| W \right\|_{2,1} + \left\| U \right\|_{2,1}) + \delta (\left\| y_s^{(1)} \oplus y_s^{(2)} ... \oplus y_s^{(V)} \right\|_F^2 \\
&+ \left\| y_s^{(1)} \otimes y_s^{(2)} ... \otimes y_s^{(V)} \ominus Y_s \right\|_F^2),
\end{aligned} \tag{12}$$

where α, β, γ and δ are trade-off parameters to keep the balance of the model. It can be observed from the objective function in Formula (12) that Y_c is confined through two terms, accounting for the correlation of the first-layer labels and the individual characteristics of the regularization. Concerning $y_s^{(i)}$, the specific label matrix is subject to three constraints. These constraints primarily focus on learning specific label matrices for all views and restrictions based on their characteristics. This splitting structure of the observed label matrix facilitates the requirement of pure variables Y_c and $y_s^{(i)}$. This, in turn, renders feature weight $W^{(i)}$ and $U^{(i)}$ more precise. Next, we propose a new methodology to transform the logic operations and relax Formula (12) to attain these variables.

Optimization

First, the objective function is not smooth due to $l_{2,1}$-norm and l_1-norm, and we relax the related terms (Nie et al. 2010) in Formula (12) including $\left\| W \right\|_{2,1} = 2Tr(W^T D W)$, $\left\| U \right\|_{2,1} = 2Tr(U^T E U)$ and $\left\| Y_n \right\|_1 = 2Tr(Y_n^T Q Y_n)$, where the diagonal matrix $d_{ii} = 1/(2 \left\| W_i \right\|_2)$, $e_{ii} = 1/(2 \left\| U_i \right\|_2)$ and $q_{ii} = 1/(2 \left| Y_n \right|)$. Then, we transform the Formula (10) as follows:

$$\min_{y_s^{(i)}} \left\| y_s^{(1)} \circ y_s^{(2)} ... \circ y_s^{(V)} \right\|_F^2, \tag{13}$$

where '$\circ$' denotes the Hadamard product. The Formula (13) represents that the result achieves minima approximately as a zero matrix, if and only if the same location in $y_s^{(i)} (i = 1, 2, \ldots, V)$ is not all labeled. Thus, Formula (13) is equal to Formula (10).

Specific to Formula (11), the key point is the same with the first term in Formula (8). The primary objective is trying to minimize the disparity between the variable generating from operations performed on various specific label matrices and the learned specific label matrix for all views. This variable is defined as Y_{tem}:

$$\begin{aligned}
&y_s^{(1)} \otimes y_s^{(2)} ... \otimes y_s^{(V)} = Y_{tem} = \sum_{i=1}^{V} y_s^{(i)} \\
&= \begin{pmatrix} y_{tem(11)} & \cdots & y_{tem(1c)} \\ \vdots & \ddots & \vdots \\ y_{tem(n1)} & \cdots & y_{tem(nc)} \end{pmatrix}. \\
&s.t. \quad \forall_{1 \le p \le n} \forall_{1 \le q \le c} y_{tem(pq)} = \begin{cases} 1 & y_{tem(pq)} \ge 1 \\ 0 & y_{tem(pq)} < 1 \end{cases}
\end{aligned} \tag{14}$$

Subsequently, we employ the multiplicative update rules to optimize the objective function, which integrate non-negative constraint conditions and introduce multipliers to restrict the variables respectively. The variable Y_s could be replaced according to Definition 1, and we replace the Frobenius norm as $\left\| Y \right\|_F^2 = Tr(Y^T Y)$. The equivalent function for Formula (12) can be written as follows:

$$\begin{aligned}
\Theta &= \sum_{i=1}^{V} Tr\left[(X^{(i)} W^{(i)} - Y_c)^T (X^{(i)} W^{(i)} - Y_c) \right] \\
&+ \sum_{i=1}^{V} Tr\left[(X^{(i)} U^{(i)} - y_s^{(i)})^T (X^{(i)} U^{(i)} - y_s^{(i)}) \right] \\
&+ \alpha Tr\left[(Y \ominus Y_c)^T (Y \ominus Y_c) \right] + 2\beta Tr(Y_n^T Q Y_n) \\
&+ 2\gamma (Tr(W^T D W) + Tr(U^T E U)) \\
&+ \delta (Tr\left[(y_s^{(1)} \circ y_s^{(2)} ... \circ y_s^{(V)})^T (y_s^{(1)} \circ y_s^{(2)} ... \circ y_s^{(V)}) \right] \\
&+ Tr[\ (Y_{tem} \ominus (Y \ominus Y_c \ominus Y_n)^T (Y_{tem} \ominus (Y \ominus Y_c \\
&\ominus Y_n)\ ]) - Tr\left(\varphi W^{(i)^T} \right) - Tr\left(\psi U^{(i)^T} \right) \\
&- Tr\left(\tau Y_c^T \right) - Tr\left(\sigma Y_n^T \right) - Tr\left(\theta y_s^{(i)^T} \right),
\end{aligned} \tag{15}$$

where $\varphi \in \mathbb{R}^{d(i) \times c}$, $\psi \in \mathbb{R}^{d(i) \times c}$, $\tau \in \mathbb{R}^{n \times c}$, $\sigma \in \mathbb{R}^{n \times c}$ and $\theta \in \mathbb{R}^{n \times c}$. To solve the non-convex problem, we fix other variables and update the remaining one in each iteration. The pseudocode is presented in Algorithm 1.

$$Y_c \leftarrow Y_c \circ \frac{A + \alpha Y + \delta B}{(V + \alpha + \delta) Y_c}, \tag{16}$$

$$Y_n \leftarrow Y_n \circ \frac{\delta C}{\beta Q Y_n + \delta Y_n}, \tag{17}$$

Algorithm 1: Double-layer Hybrid-Label Identification

Input: Data matrices $\left\{\boldsymbol{X^{(i)}}\right\}_{\boldsymbol{i=1}}^{\boldsymbol{V}}$; Label matrix $\boldsymbol{Y}$.
Parameter: Parameters α, β, γ and δ.
Output: Set of selected features.

Initialize $\boldsymbol{Y_c}$, $\boldsymbol{Y_n}$, $\boldsymbol{W^{(i)}}$, $\boldsymbol{U^{(i)}}$, $\boldsymbol{y_s^{(i)}}$;
repeat
  Update the matrix $\boldsymbol{Y_c}$ according to Formula (16);
  Update the matrix $\boldsymbol{Y_n}$ according to Formula (17);
  Update the matrix $\boldsymbol{W^{(i)}}$ according to Formula (18);
  Update the matrix $\boldsymbol{U^{(i)}}$ according to Formula (19);
  Update the matrix $\boldsymbol{y_s^{(i)}}$ according to Formula (20);
  Update the objective function (15);
until Convergence
Obtain the ordered feature sequence by calculating $||(\boldsymbol{W}+\boldsymbol{U})_{(j)}||_2$ where $j = 1, 2, 3, ..., d$;
return Top ranked features as ***s-DHLI-f***.

$$W^{(i)} \leftarrow W^{(i)} \circ \frac{{X^{(i)}}^T Y_c}{{X^{(i)}}^T X^{(i)} W^{(i)} + \gamma D^{(i)} W^{(i)}}, \tag{18}$$

$$U^{(i)} \leftarrow U^{(i)} \circ \frac{{X^{(i)}}^T y_s^{(i)}}{{X^{(i)}}^T X^{(i)} U^{(i)} + \gamma E^{(i)} U^{(i)}}, \tag{19}$$

$$y_s^{(i)} \leftarrow y_s^{(i)} \circ \frac{X^{(i)} U^{(i)} + \delta H}{(1+\delta) y_s^{(i)} + \delta y_s^{(i)} \circ F \circ F}, \tag{20}$$

where $A = \sum_{i=1}^{V} X^{(i)} W^{(i)}$, $B = (Y \ominus Y_n \ominus Y_{tem})$, $C = (Y \ominus Y_c \ominus Y_{tem})$, $F = y_s^{(1)} \circ y_s^{(2)} ... \circ y_s^{(i-1)} \circ y_s^{(i+1)} ... \circ y_s^{(V)}$, $G = y_s^{(1)} \otimes y_s^{(2)} ... \otimes y_s^{(i-1)} \otimes y_s^{(i+1)} ... \otimes y_s^{(V)}$ and $H = (Y \ominus Y_n \ominus Y_c \ominus G)$.

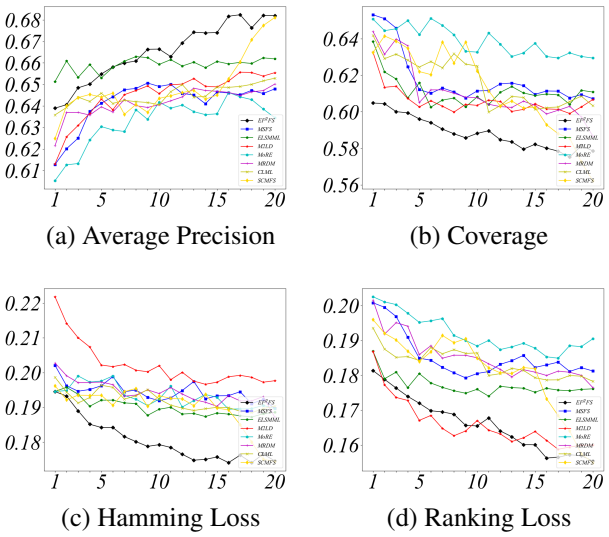
Complexity Analysis

The optimization procedure consists of five primary components, and for simplicity, we assume that the dimension for the feature matrix of each view is d. Specifically, when updating Y_c, the computation complexity is $O(ndc)$, while the computation complexity for updating Y_n, $W^{(i)}$, $U^{(i)}$ and $y_s^{(i)}$ are $O(nc)$, $O(nd^2c)$, $O(nd^2c)$ and $O(ndc)$, respectively. As such, the whole training procedure can be conservatively approximated as $O(nd^2c)$ for each iteration.

Experiments

Experimental Setup

Datasets. Following (Zhu, Li, and Zhang 2015; Zhang et al. 2020b; Li and Chen 2021), six popular multi-view multi-label datasets are adopted to facilitate a fair result comparison with state-of-the-art methods, including SCENE, OBJECT, MIRFlickr, Corel5K, IAPRTC12 and 3Sources. Table 1 illustrates the characteristics of the datasets in the experiments. The views are identified by their individual names, and the number of features associated with each view are indicated within braces. In situations where a view is absent, it is denoted by '-'. The total number of samples, features and labels is denoted as n, d and c, respectively.



(a) Average Precision (b) Coverage

(c) Hamming Loss (d) Ranking Loss

Figure 2: Eight methods on MIRFlickr in terms of Average Precision, Coverage, Hamming Loss and Ranking loss.

Comparing Methods. We implement seven embedded methods that focus on the feature space, including three multi-view multi-label learning methods (MSFS (Zhang et al. 2020b), ELSMML (Liu et al. 2023a) and M2LD (Liu et al. 2023c)) and four multi-label learning methods (MoRE (Liu et al. 2022), MRDM (Huang and Wu 2021), CLML (Li et al. 2022) and SCMFS (Hu et al. 2020)). Furthermore, the learning model used allowed for the extraction of a matrix that reflects the significance of each feature.

Evaluation Metrics. On each dataset, five-fold cross-validation is performed and the mean accuracy, as well as standard deviation, are reported. The parameters of the regularization paradigm are searched in set $\left\{10^{-3}, 10^{-2}, ..., 10^{3}\right\}$, and four metrics commonly used in this field are selected to evaluate these methods. i.e., Average Precision (AP), Coverage, Hamming Loss (HL) and Ranking Loss (RL). For the first metric, the value is larger with better performance, and the following three metrics are opposite. The detailed explanation about the metrics is introduced in (Zhang and Zhou 2013; Gibaja and Ventura 2015).

Experimental Results

Feature Selection Performance. Table 2-3 list the performance results of different methods for varying percentage of features, with the range of percentages from one to twenty. It can be observed that: (1) DHLI achieves a bright performance, outperforming other methods in most cases, except that DHLI and ELSMML have the same ranking

Feature views	SCENE	OBJECT	MIRFlickr	Corel5K	IAPRTC12	3Sources
View1(d_1)	CH(64)	CH(64)	DH(100)	DH(100)	DH(100)	BBC(1000)
View2(d_2)	CM(225)	CM(225)	GIST(512)	DHV3H1(300)	DHV3H1(300)	Reuters(1000)
View3(d_3)	CORR(144)	CORR(144)	HH(100)	GIST(512)	GIST(512)	Guardian(1000)
View4(d_4)	EDH(73)	EDH(73)	-	HHV3H1(300)	HHV3H1(300)	-
View5(d_5)	WT(128)	WT(128)	-	HH(100)	HH(100)	-
Samples(n)	4400	6047	4053	4999	4999	169
Features(d)	634	634	712	1312	1312	3000
Labels(c)	33	31	38	260	260	6

Table 1: Characteristics of the datasets in our experiments.

Dataset	DHLI	MSFS	ELSMML	M2LD	MoRE	MRDM	CLML	SCMFS
				AP ↑				
SCENE	**0.7951±0.012**	0.7887±0.008	0.7801±0.01	0.7792±0.014	0.7877±0.007	0.7929±0.01	0.787±0.008	0.7866±0.008
OBJECT	**0.4845±0.046**	0.4626±0.029	0.4632±0.021	0.4531±0.026	0.4719±0.031	0.4595±0.02	0.4748±0.024	0.4557±0.019
MIRFlickr	**0.6635±0.015**	0.6417±0.01	0.6587±0.006	0.6436±0.013	0.6322±0.011	0.6407±0.008	0.6442±0.007	0.6471±0.013
Corel5K	**0.237±0.009**	0.1861±0.007	0.2317±0.005	0.2355±0.008	0.1845±0.023	0.2339±0.011	0.233±0.008	0.1801±0.01
IAPRTC12	**0.1421±0.007**	0.1385±0.005	0.1399±0.006	0.137±0.006	0.1263±0.006	0.14±0.005	0.1352±0.002	0.1301±0.004
3Sources	**0.4534±0.034**	0.4252±0.035	0.4378±0.04	0.4368±0.038	-	0.4196±0.041	0.4232±0.046	0.4241±0.044
				Coverage ↓				
SCENE	**0.4174±0.017**	0.4271±0.012	0.4569±0.016	0.4416±0.016	0.4295±0.01	0.4252±0.017	0.4308±0.012	0.4318±0.011
OBJECT	**0.2979±0.026**	0.3005±0.027	0.3071±0.017	0.3142±0.03	0.302±0.03	0.309±0.031	0.298±0.02	0.3106±0.016
MIRFlickr	**0.5897±0.013**	0.6173±0.013	0.6124±0.011	0.6069±0.009	0.6397±0.01	0.6128±0.017	0.6191±0.015	0.6136±0.021
Corel5K	**0.4832±0.011**	0.5085±0.012	0.4898±0.007	0.508±0.008	0.5378±0.012	0.4918±0.021	0.4951±0.004	0.5357±0.025
IAPRTC12	**0.5091±0.009**	0.5098±0.009	0.5094±0.007	0.5131±0.011	0.5877±0.02	0.5097±0.017	0.5149±0.005	0.6264±0.014
3Sources	**0.6181±0.029**	0.6602±0.037	**0.6181±0.023**	0.6397±0.03	-	0.6723±0.042	0.6727±0.046	0.6716±0.047

Table 2: Experimental results (mean ± std) in terms of Average Precision and Coverage, where the 1st/2nd best results are shown in boldface/underline.

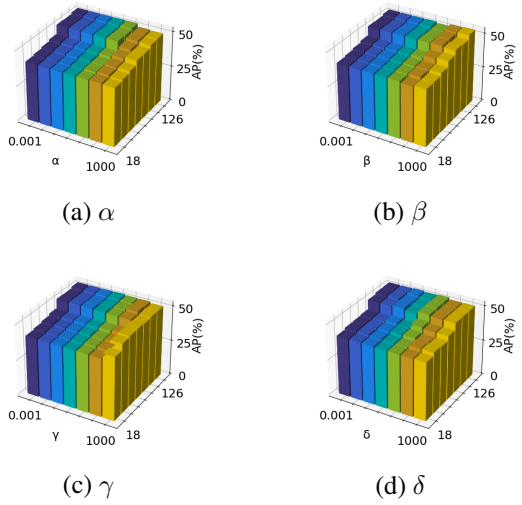
jointly for the Coverage metric on the 3Sources dataset. (2) DHLI achieves superior performance against MSFS, M2LD, MoRE, CLML and SCMFS, which seldom rank first or second. ELSMML and MRDM perform well and rank first or second in 41.7% and 29.2% of cases, respectively. This is because they comply more accurately with the constraint of the label matrix. (3) In the context of multi-view multi-label methods, the input matrix of features is typically smaller compared to that of multi-label methods. As a result, these methods require relatively less memory space during experimental procedures. Therefore, it is actually beneficial for dealing with large-scale semantic content. We also illustrate one dataset for all metrics in Figure 2 to show our performance clearly.

Parameter Analysis. In DHLI method, there are four trade-off parameters α, β, γ and δ that influence the performance results. Figure 3 shows how four parameters affect the Average precision on the OBJECT dataset. The parameter is individually tuned while keeping the other parameters fixed, and the grid search is conducted over a predefined range. Initially, the results typically exhibit an incline, later stabilizing as the number of selected features increases. Nonetheless, due to divergent parameter configurations, insubstantial discrepancies may surface under the same number of selected features, eliciting unwarranted variance in results. Hence, we can infer the proposed method relative stability solely with a substantial number of instances.

Ablation Study. To evaluate the effectiveness for each part of the objective function, we perform ablation experiments on SCENE, Corel5K and IAPRTC12. The different functions consisted of partial DHLI, are presented in Table 4 that DHLI-alpha remains the regularization for the assumption of the majority in the common part. DHLI-beta preserves the constraints for the sparsity of the noises in the label matrix. DHLI-gamma guarantees the sparsity of the feature matrix on rows, and DHLI-delta ensures the characteristics of specific labels for each view. These findings reveal that the removal of any part has been observed to have a detrimental impact on the results of Average Precision, emphasizing the significance of all constituent parts. Notably, the absence of constraints for outstanding the specific causes the most significant decrease in performance compared to other missing parts. Therefore, it can be inferred that the contribution of each component is indispensable for attaining the best performance of the model.

Dataset	DHLI	MSFS	ELSMML	M2LD	MoRE	MRDM	CLML	SCMFS
				HL ↓				
SCENE	**0.0969±0.006**	0.0999±0.005	0.1068±0.004	0.104±0.006	0.1007±0.005	0.0979±0.005	0.101±0.005	0.1003±0.005
OBJECT	**0.0572±0.004**	0.0574±0.003	0.0597±0.002	0.0597±0.003	0.0576±0.003	0.0583±0.003	0.0573±0.002	0.0603±0.002
MIRFlickr	**0.1808±0.007**	0.1944±0.002	0.1901±0.003	0.2029±0.007	0.1931±0.003	0.1946±0.003	0.1917±0.003	0.1901±0.006
Corel5K	**0.01379±0.00**	0.01381±0.00	0.0138±0.00	0.01465±0.00	0.01381±0.00	0.0138±0.00	0.01382±0.00	0.01427±0.00
IAPRTC12	**0.01549±0.00**	0.01561±0.00	0.01551±0.00	0.01653±0.00	0.0155±0.00	0.0156±0.00	0.01554±0.00	0.01993±0.00
3Sources	**0.2315±0.015**	0.2357±0.017	0.2331±0.013	0.2491±0.025	-	0.2358±0.017	0.2471±0.018	0.2454±0.017
				RL ↓				
SCENE	**0.0926±0.008**	0.097±0.006	0.1041±0.007	0.1037±0.009	0.0974±0.005	0.0949±0.007	0.098±0.006	0.0985±0.005
OBJECT	**0.1642±0.021**	0.1648±0.017	0.1756±0.012	0.1791±0.019	0.1661±0.016	0.1744±0.017	0.1658±0.013	0.1777±0.012
MIRFlickr	**0.1671±0.009**	0.1855±0.006	0.1775±0.004	0.1672±0.008	0.1925±0.006	0.1858±0.007	0.1844±0.006	0.182±0.011
Corel5K	**0.2283±0.007**	0.2665±0.009	0.232±0.005	0.2379±0.005	0.2613±0.008	0.2343±0.015	0.2357±0.003	0.263±0.02
IAPRTC12	**0.2068±0.011**	0.2093±0.006	0.2087±0.006	0.207±0.009	0.2452±0.01	0.2069±0.013	0.2134±0.003	0.2402±0.012
3Sources	**0.4761±0.035**	0.5253±0.043	0.4875±0.045	0.5045±0.039	-	0.536±0.051	0.5345±0.055	0.5347±0.053

Table 3: Experimental results (mean ± std) in terms of Hamming Loss and Ranking Loss, where the 1st/2nd best results are shown in boldface/underline.

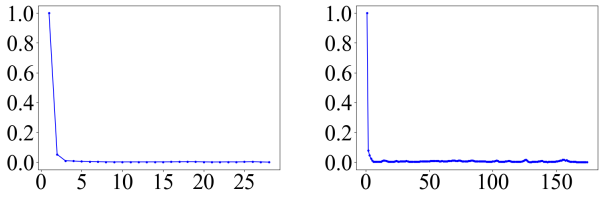


(a) α (b) β (c) γ (d) δ

Figure 3: Parameter sensitivity studies on the OBJECT dataset.

DHLI-alpha	DHLI-beta	DHLI-gamma	DHLI-delta	SCENE	Corel5K	IAPRTC12
	√	√	√	0.7852	0.2366	0.1403
√		√	√	0.7851	0.2369	0.1401
√	√		√	0.7853	0.2364	0.1396
√	√	√		0.7842	0.2316	0.1390
√	√	√	√	0.7951	0.2370	0.1421

Table 4: Ablation experimental results of DHLI on SCENE, Corel5K and IAPRTC12.



(a) OBJECT (b) MIRFlickr

Figure 4: Convergence curves on OBJECT and MIRFlickr.

Convergence. Figure 4 shows the convergence of LGCM on OBJECT and MIRFlickr. To facilitate observation of the oscillation, we set the initial point to approximately one. The chosen stopping criterion is $\left|z^t - z^{t-1}\right| / z^{t-1} < \varepsilon$. X-axis represents the number of iterations, and Y-axis is similar to the stopping criterion except absolute. It can be seen clearly that the optimization process converges quickly.

Conclusion

This paper offers a novel investigation into the splitting structure of observed labels for MVML. It also provides a thorough analysis of the relationship among hybrid labels in a double layer, which comprehensively considers the attributes of MVML, including the common and view-specific components. Additionally, a pure mapping between labels and feature representations is constructed with the aim of accurately determining feature weights. The proposed method, DHLI, is demonstrated to outperform other state-of-the-art approaches via experiments in MVML. As such, this research provides valuable insights that we hope will inspire further investigation in this direction, and utilize more powerful techniques to increase interpretability of our method.

Acknowledgments

This work is funded by: by Science Foundation of Jilin Province of China under Grant No. 20230508179RC, and China Postdoctoral Science Foundation funded project under Grant No. 2023M731281, and Changchun Science and Technology Bureau Project 23YQ05.

References

Chang, X.; and Yang, Y. 2016. Semisupervised feature analysis by mining correlations among multiple tasks. *IEEE transactions on neural networks and learning systems*, 28(10): 2294–2305.

Cui, L.; Bai, L.; Wang, Y.; Philip, S. Y.; and Hancock, E. R. 2021. Fused lasso for feature selection using structural information. *Pattern Recognition*, 119: 108058.

Efron, B.; Hastie, T.; Johnstone, I.; and Tibshirani, R. 2004. Least angle regression.

Eswaran, D.; Kumar, S.; and Faloutsos, C. 2020. Higher-order label homogeneity and spreading in graphs. In *Proceedings of The Web Conference 2020*, 2493–2499.

Fan, Y.; Liu, J.; Weng, W.; Chen, B.; Chen, Y.; and Wu, S. 2021. Multi-label feature selection with local discriminant model and label correlations. *Neurocomputing*, 442: 98–115.

Fu, K.; Du, C.; Wang, S.; and He, H. 2022. Multi-View Multi-Label Fine-Grained Emotion Decoding From Human Brain Activity. *IEEE Transactions on Neural Networks and Learning Systems*.

Gibaja, E.; and Ventura, S. 2015. A tutorial on multilabel learning. *ACM Computing Surveys (CSUR)*, 47(3): 1–38.

Gong, X.; Yuan, D.; and Bao, W. 2021. Understanding partial multi-label learning via mutual information. *Advances in Neural Information Processing Systems*, 34: 4147–4156.

González-López, J.; Ventura, S.; and Cano, A. 2019. Distributed selection of continuous features in multilabel classification using mutual information. *IEEE transactions on neural networks and learning systems*, 31(7): 2280–2293.

Hu, L.; Li, Y.; Gao, W.; Zhang, P.; and Hu, J. 2020. Multi-label feature selection with shared common mode. *Pattern Recognition*, 104: 107344.

Huang, J.; Li, G.; Huang, Q.; and Wu, X. 2016. Learning label-specific features and class-dependent labels for multi-label classification. *IEEE transactions on knowledge and data engineering*, 28(12): 3309–3323.

Huang, R.; and Wu, Z. 2021. Multi-label feature selection via manifold regularization and dependence maximization. *Pattern Recognition*, 120: 108149.

Jian, L.; Li, J.; Shu, K.; and Liu, H. 2016. Multi-label informed feature selection. In *IJCAI*, volume 16, 1627–33.

Lee, J.; and Kim, D.-W. 2015. Mutual information-based multi-label feature selection using interaction information. *Expert Systems with Applications*, 42(4): 2013–2025.

Lee, J.; and Kim, D.-W. 2017. SCLS: Multi-label feature selection based on scalable criterion for large label set. *Pattern Recognition*, 66: 342–352.

Li, J.; Li, P.; Hu, X.; and Yu, K. 2022. Learning common and label-specific features for multi-Label classification with correlation information. *Pattern Recognition*, 121: 108259.

Li, X.; and Chen, S. 2021. A concise yet effective model for non-aligned incomplete multi-view and missing multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 5918–5932.

Lim, H.; and Kim, D.-W. 2020. MFC: Initialization method for multi-label feature selection based on conditional mutual information. *Neurocomputing*, 382: 40–51.

Lin, Y.; Hu, Q.; Liu, J.; and Duan, J. 2015. Multi-label feature selection based on max-dependency and min-redundancy. *Neurocomputing*, 168: 92–103.

Lin, Y.; Hu, Q.; Liu, J.; Zhu, X.; and Wu, X. 2021. MULFE: multi-label learning via label-specific feature space ensemble. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(1): 1–24.

Lin, Y.; Liu, H.; Zhao, H.; Hu, Q.; Zhu, X.; and Wu, X. 2022. Hierarchical feature selection based on label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*.

Liu, B.; Li, W.; Xiao, Y.; Chen, X.; Liu, L.; Liu, C.; Wang, K.; and Sun, P. 2023a. Multi-view multi-label learning with high-order label correlation. *Information Sciences*, 624: 165–184.

Liu, C.; Wen, J.; Luo, X.; Huang, C.; Wu, Z.; and Xu, Y. 2023b. DICNet: Deep Instance-Level Contrastive Network for Double Incomplete Multi-View Multi-Label Classification. *arXiv preprint arXiv:2303.08358*.

Liu, M.; Luo, Y.; Tao, D.; Xu, C.; and Wen, Y. 2015. Low-rank multi-view learning in matrix completion for multi-label image classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Liu, S.; Song, X.; Ma, Z.; Ganaa, E. D.; and Shen, X. 2022. MoRE: multi-output residual embedding for multi-label classification. *Pattern Recognition*, 126: 108584.

Liu, W.; Yuan, J.; Lyu, G.; and Feng, S. 2023c. Label driven latent subspace learning for multi-view multi-label classification. *Applied Intelligence*, 53(4): 3850–3863.

Luo, M.; Chang, X.; Nie, L.; Yang, Y.; Hauptmann, A. G.; and Zheng, Q. 2017. An adaptive semisupervised feature analysis for video semantic recognition. *IEEE transactions on cybernetics*, 48(2): 648–660.

Luo, Y.; Tao, D.; Xu, C.; Xu, C.; Liu, H.; and Wen, Y. 2013. Multiview vector-valued manifold regularization for multilabel image classification. *IEEE transactions on neural networks and learning systems*, 24(5): 709–722.

Lyu, G.; Deng, X.; Wu, Y.; and Feng, S. 2022. Beyond shared subspace: A view-specific fusion for multi-view multi-label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 7647–7654.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. 2010. Efficient and robust feature selection via joint l2, 1-norms minimization. *Advances in neural information processing systems*, 23.

Qian, W.; Xiong, C.; Qian, Y.; and Wang, Y. 2022. Label enhancement-based feature selection via fuzzy neighborhood discrimination index. *Knowledge-Based Systems*, 250: 109119.

Tan, Q.; Yu, G.; Wang, J.; Domeniconi, C.; and Zhang, X. 2019. Individuality-and commonality-based multiview multilabel learning. *IEEE transactions on cybernetics*, 51(3): 1716–1727.

Wu, X.; Chen, Q.-G.; Hu, Y.; Wang, D.; Chang, X.; Wang, X.; and Zhang, M.-L. 2019. Multi-View Multi-Label Learning with View-Specific Information Extraction. In *IJCAI*, 3884–3890.

Wu, X.; Jiang, B.; Yu, K.; Chen, H.; and Miao, C. 2020. Multi-label causal feature selection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 6430–6437.

Xiong, C.; Qian, W.; Wang, Y.; and Huang, J. 2021. Feature selection based on label distribution and fuzzy mutual information. *Information Sciences*, 574: 297–319.

Yin, J.; and Zhang, W. 2023. Multi-view multi-label learning with double orders manifold preserving. *Applied Intelligence*, 53(12): 14703–14716.

Yu, G.; Chen, X.; Domeniconi, C.; Wang, J.; Li, Z.; Zhang, Z.; and Wu, X. 2018. Feature-induced partial multi-label learning. In *2018 IEEE international conference on data mining (ICDM)*, 1398–1403. IEEE.

Yu, K.; Cai, M.; Wu, X.; Liu, L.; and Li, J. 2021. Multilabel feature selection: a local causal structure learning approach. *IEEE Transactions on Neural Networks and Learning Systems*.

Yuan, M.; and Lin, Y. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1): 49–67.

Zhang, C.; Yu, Z.; Hu, Q.; Zhu, P.; Liu, X.; and Wang, X. 2018. Latent semantic aware multi-view multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhang, J.; Lin, Y.; Jiang, M.; Li, S.; Tang, Y.; and Tan, K. C. 2020a. Multi-label Feature Selection via Global Relevance and Redundancy Optimization. In *IJCAI*, 2512–2518.

Zhang, J.; Luo, Z.; Li, C.; Zhou, C.; and Li, S. 2019. Manifold regularized discriminative feature selection for multi-label learning. *Pattern Recognition*, 95: 136–150.

Zhang, M.-L.; and Zhang, K. 2010. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 999–1008.

Zhang, M.-L.; and Zhou, Z.-H. 2013. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8): 1819–1837.

Zhang, P.; Liu, G.; and Gao, W. 2019. Distinguishing two types of labels for multi-label feature selection. *Pattern Recognition*, 95: 72–82.

Zhang, Y.; Wu, J.; Cai, Z.; and Philip, S. Y. 2020b. Multi-view multi-label learning with sparse feature selection for image annotation. *IEEE Transactions on Multimedia*, 22(11): 2844–2857.

Zhao, D.; Gao, Q.; Lu, Y.; and Sun, D. 2022a. Learning view-specific labels and label-feature dependence maximization for multi-view multi-label classification. *Applied Soft Computing*, 124: 109071.

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

Zhu, C.; Miao, D.; Wang, Z.; Zhou, R.; Wei, L.; and Zhang, X. 2020. Global and local multi-view multi-label learning. *Neurocomputing*, 371: 67–77.

Zhu, X.; Huang, Z.; Cheng, H.; Cui, J.; and Shen, H. T. 2013. Sparse hashing for fast multimedia search. *ACM Transactions on Information Systems (TOIS)*, 31(2): 1–24.

Zhu, X.; Li, X.; and Zhang, S. 2015. Block-row sparse multiview multilabel learning for image classification. *IEEE transactions on cybernetics*, 46(2): 450–461.