Semi-Supervised Multi-View Multi-Label Learning with View-Specific Transformer and Enhanced Pseudo-Label

Quanjiang Li[1], Tingjin Luo[1]*, Mingdie Jiang[1], Zhangqi Jiang[1], Chenping Hou[1], Feijiang Li[2]

[1]College of Science, National University of Defense Technology, Changsha 410073, Hunan, China
[2]Institute of Big Data Science and Industry, Shanxi University, Taiyuan, 030006, Shanxi, China
liquanjiang@nudt.edu.cn, tingjinluo@hotmail.com, jiangmingdie20@nudt.edu.cn, jiangzq@nudt.edu.cn, hcpnudt@hotmail.com, fjli@sxu.edu.cn

Abstract

Multi-view multi-label learning has become a research focus for describing objects with rich expressions and annotations. However, real-world data often contains numerous unlabeled instances, due to the high cost and technical limitations of manual labeling. This crucial problem involves three main challenges: i) How to extract advanced semantics from available views? ii) How to build a refined classification framework with limited labeled space? iii) How to provide more high-quality supervisory information? To address these problems, we propose a Semi-Supervised Multi-View Multi-Label Learning Method with View-Specific Transformer and Enhanced Pseudo-Label named SMVTEP. Specifically, Generative Adversarial Networks are employed to extract informative shared and specific representations and their consistency and distinctiveness are ensured through the adversarial mechanism and information theory based contrastive learning. Then we build specific classifiers for each extracted feature and apply instance-level manifold constraints to reduce bias across classifiers. Moreover, we design a transformer-style fusion approach that simultaneously captures the imbalance of expressive power among views, mapping effects on specific labels, and label dependencies by incorporating confidence scores and category semantics into the self-attention mechanism. Furthermore, after using Mixup for data augmentation, category-enhanced pseudo-labels are leveraged to improve the reliability of additional annotations by aligning the label distribution of unlabeled samples with the true distribution. Finally, extensive experimental results validate the effectiveness of SMVTEP against state-of-the-art methods.

Introduction

Multi-label learning has garnered increasing attention due to its broad applications in text classification (Chang et al. 2020), image annotation (Sun and Xie 2024), and computer vision tasks (Jung et al. 2024). Since the proliferation of data sources and advancements in feature extraction methods, multi-view data has become widely available and provides greater opportunities for comprehensive descriptions of observed targets (Hu et al. 2023; Guan et al. 2024; Hu et al. 2024a). For example, autonomous vehicles integrate data from cameras, LiDAR, and radar to enhance environmental perception and driving safety (Xin et al. 2023). Besides, multi-view data encompasses abundant perspectives (Hu et al. 2024b), which can be combined with multi-label to convey the rich semantic structure of complex data (Xiao et al. 2024). Therefore, this paper focuses on the multi-view multi-label classification (MvMLC) task.

For traditional MvMLC methods, they assume that the given data is complete, which is often violated in practice (Zhu et al. 2023). Due to the limitation of technical capability and resource cost, the acquisition of perfectly labeled data is not light-hearted. To handle the issue of insufficient labels under multi-view, MVEL-ILD (Zhang et al. 2013) utilized label correlations to achieve consistent outcomes, while DD-IMvMLC (Wen et al. 2023) introduced a missing indicator matrix to focus on the available label information. These MvMLC methods under incomplete data mainly address the partial absence of multiple labels, leveraging the available subset to infer the missing ones. However, in reality, multi-view data tends to suffer from unlabeled instances. For example, in the healthcare diagnostics (Miller and Lowengrub 2022), factors such as resource availability, patient condition diversity, and diagnostic complexity can result in many records remaining unlabeled, potentially decreasing the diagnostic accuracy and reliability. The numerous views and label categories, coupled with a lack of sufficient labeled samples, severely hinder the information demands of model deployment and inflict a significant blow to MvMLC methods. Therefore, few methods available today can well address the tricky issue. In fact, solving the semi-supervised problem primarily involves utilizing large amounts of unsupervised data and extracting substantial supervised information (Sun et al. 2021). For MvMLC, the goal is to obtain advanced feature representations that align with label semantics, establish fine-grained associations between features and specific labels in the limited annotation space, and uncover internal correlations between label categories and additional annotations with minimal noise.

To tackle these problems, we propose a semi-supervised multi-view multi-label learning method with view-specific transformer and enhanced pseudo-label named SMVTEP. The motivation for SMVTEP arises from the following aspects: 1) The emphasis of multi-view learning is on capturing shared information across various views while pre-

*Corresponding author.

serving the unique information of each individual view (Sun and Wang 2024). Therefore, Generative Adversarial Networks are employed to extract clear and informative representations. Besides, we utilize the adversarial mechanism for achieving consistency in the shared features and apply the contrastive learning based on the information theory to limit feature interactions and maintain the distinctiveness of specific representations. 2) Each view fails to predict all labels accurately and excels at distinguishing a particular subset (Han et al. 2024). Then we consider building specific classifiers for each extracted feature and apply manifold constraints grounded in the structural similarity between features and multi-label predictions to reduce bias of different classifiers. 3) During the process of result fusion, internal correlations within both the feature and label spaces and the mapping semantics are crucial factors. Thus, a transformer guided by both confidence scores and true label semantics is utilized to capture the relative importance between the shared and specific views, the significance of each view for specific subcategory and the dependencies among labels. 4) Due to the scarcity of labeled samples, the Mixup algorithm is used for data augmentation to prevent the model from overfitting. Subsequently, pseudo-labels are assigned to explore additional supervisory information. However, distribution imbalance in multi-label causes dominant labels prevailing in pseudo-labels and leads to accumulated noise (Su and Xu 2024). To this end, category-enhanced pseudo-labels are leveraged to ensure that the label proportions of each class in the unlabeled samples match those in the labeled samples, thereby making the pseudo-label distribution as close to the true distribution as possible and enhancing the reliability. It is no doubt that our SMVTEP not only extracts rich feature representations and deep semantic associations in the limited annotation environment, but also stably provides high-quality supervisory information, making it highly suitable for multi-view semi-supervised multi-label problems. Our contributions are summarized as follows:

- To our knowledge, this is the first deep classification model grounded in the application of Generative Adversarial Networks and Transformer, capable of handling few labeled instances in multi-view multi-label data.
- SMVTEP is a unified framework designed to obtain semantically distinct high-level feature representations, develop discriminative classification techniques, implement comprehensive late-stage fusion mechanisms, and extract reliable supervisory information.
- Extensive experimental results demonstrate that our SMVTEP consistently outperforms other approaches and confirm its effectiveness and robustness.

Related Work

Incomplete Multi-View Multi-Label Learning

Addressing incomplete data in MvMLC initially relied on traditional approaches. iMVWL (Tan et al. 2018) leveraged weak label correlations to mitigate the impact of missing labels. TM3L (Zhao et al. 2021) combined label matrix completion with kernel extreme learning machines. NAIM3L (Li and Chen 2021) exploited both consensus across diverse views and the global and local structures among multiple labels from rank constraint. However, these shallow models struggled with capturing complex feature semantics and label correlations. Recently, deep learning based methods have shown remarkable performance. DICNet (Liu et al. 2023a) built a feature extraction framework and introduced the instance-level contrastive learning for consensus representations. LMVCAT (Liu et al. 2023b) aggregated features using transformer-based modules and handled missing labels with label-guided graph constraints, while VIST (Ou et al. 2024) implemented view-category interactive sharing transformers and category consistency guided embedding module to improve discriminative power.

Semi-Supervised Multi-Label Learning

Semi-supervised multi-label learning is typically categorized into inductive and transductive types. Inductive learning purely relies on model training to predict test data. Coins (Zhan and Zhang 2017) used co-training with feature space splitting and pairwise ranking on unlabeled data to optimize multi-label classification. SCTML (Li et al. 2024) designed a two-layer stacking framework that captured label correlations in both base and meta learners, integrating co-training and manifold assumptions. Transductive learning assumes test data comes from unlabeled data and propagates labels based on sample similarity. SSWL (Dong, Li, and Zhou 2018) applied the integration of multiple models to optimize classifiers. MSWL (Zhang et al. 2020) employed a manifold regularization sparse model to learn from unlabeled data. ESMC (Akbarnejad and Baghshah 2018) employed pseudo-instance parameterized sparse Gaussian models to effectively leverage unlabeled data. MGLP (Hu and Miao 2022) utilized a three-way decision method to select unlabeled data for further annotation.

Methodology

In this section, we will explain our SMVTEP through the following four aspects shown in Fig. 1: shared and specific information extraction, consistent view-specific label learning, adaptively label-guided transformer fusion and category-enhanced pseudo-label strategy.

Notations and Problem Formulation

The multi-view dataset is defined with N instances and V views as $\mathcal{X} = \{\boldsymbol{X}^{(v)}\}_{v=1}^{V}$, where $\boldsymbol{X}^{(v)} = \{\boldsymbol{x}_i^{(v)}\}_{i=1}^{N} \in \mathbb{R}^{N \times d_v}$ is the d_v-dimensional feature matrix of the v-th view. We let $\boldsymbol{Y} \in \{0,1\}^{N \times C}$ represent the label matrix, where C is the number of categories. Besides, $\boldsymbol{Y}_i \in \{0,1\}^C$ is a label vector and $Y_{i,j} = 1$ if the sample i belongs to class j, otherwise $Y_{i,j} = 0$. The number of labeled and unlabeled instances is n_l and n_u, which satisfy that $n_l \ll N$ and $n_l + n_u = N$. We also denote $\mathcal{F}$ and $\mathcal{R}$ as the index spaces for labeled and unlabeled instances, respectively.

Shared and Specific Representation Extraction

Learning an expressive representation from multi-view data is essential to reduce redundancy and noise in raw data, as

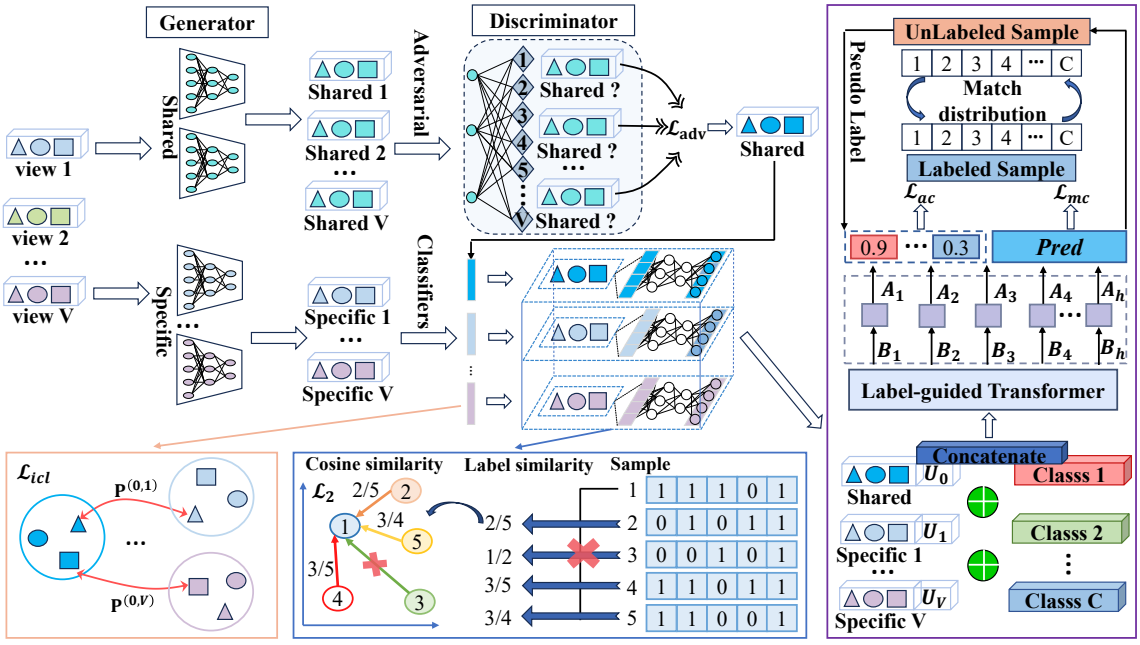


Figure 1: The main framework of our proposed SMVTEP. Different shapes signify different samples.

well as promote information integration (Jia et al. 2020). Therefore, both common semantic and specific discriminative representations should be fully exploited to enhance information utilization efficiency. To ensure consistency in the underlying representations of the same sample across multiple views and maintain the unique contribution of each view, we employ Generative Adversarial Networks (Wu et al. 2019) to learn the shared and specific subspace. The generators G and R are employed to extract the common and specific features into the same dimension M, i.e., the corresponding representation from the v-th view is $\boldsymbol{h}_c^v = G_v(\boldsymbol{x}^v)$ and $\boldsymbol{h}_s^v = R_v(\boldsymbol{x}^v)$. The discriminator D is a V-classes classifier designed to determine the original view from which the generated shared information is derived. Let $\boldsymbol{z}_i^v$ be a V-dimensional one-hot vector that indicates the source of $\boldsymbol{h}_{c,i}^v$, where only the v-th element equals to 1 and all other elements are 0. Denoting the output $\hat{\boldsymbol{z}}_i^v = D\left(\boldsymbol{h}_{c,i}^v\right)$, the adversarial loss can be utilized to confuse the discriminator:

$$\mathcal{L}_{adv} = \mathcal{F}\left(-\sum_{i=1}^{N}\sum_{v=1}^{V}\boldsymbol{z}_i^v \log \hat{\boldsymbol{z}}_i^v\right), \tag{1}$$

where $\mathcal{F}(\cdot)$ is a monotonically decreasing function and we set $\mathcal{F}(x) = e^{-x}$. In this way, the discriminator is unable to recognize the true view of the shared representations, which implies that the shared information is consistent across different views. Hence, we obtain the shared feature by $\boldsymbol{h}_c = \frac{1}{V}\sum_{l=1}^{V}\boldsymbol{h}_c^l$. To minimize the overlap between the shared and specific information, most methods (Wu et al. 2019; Jia et al. 2020) employ the constraint of geometric orthogonality. However, the interactions between different views are undoubtedly more complex than simple linear geometric relationships. Such constraint may leave abundant redundant information during feature extraction. Therefore, we employ the contrastive learning guided by the information theory (Lin et al. 2022) to directly constrain their mutual information, aiming to stably limit intricate interactions. To calculate the mutual information, the Softmax activation function σ_s is used to form the distribution probability vectors $\widetilde{\boldsymbol{h}}_{c,i}$ and $\widetilde{\boldsymbol{h}}_{s,i}^v$ (Peng et al. 2019). In other words, $\widetilde{\boldsymbol{h}}_c$ and $\widetilde{\boldsymbol{h}}_s^v$ can be interpreted as two discrete cluster assignment variables over M categories. Then the joint probability distribution can be computed as below:

$$\boldsymbol{P}^{(v,c)} = \sum_{i=1}^{N}\left(\widetilde{\boldsymbol{h}}_{s,i}^v\right)^T \widetilde{\boldsymbol{h}}_{c,i}. \tag{2}$$

The contrastive loss between the shared and specific representations is defined as $\ell_{v,c} = I\left(\widetilde{\boldsymbol{h}}_c; \widetilde{\boldsymbol{h}}_s^v\right) + \alpha\left(H\left(\widetilde{\boldsymbol{h}}_c\right) + H\left(\widetilde{\boldsymbol{h}}_s^v\right)\right)$, where I and H denote the mutual information and entropy. From the loss, a smaller $I\left(\widetilde{\boldsymbol{h}}_c; \widetilde{\boldsymbol{h}}_s^v\right)$ signifies less consistent interaction and a smaller $H\left(\widetilde{\boldsymbol{h}}_c\right)$ and $H\left(\widetilde{\boldsymbol{h}}_s^v\right)$ indicates the representations carry more effective information and have lower uncertainty. By denoting the marginal probability distribution as $\boldsymbol{P}^{(v)}$ and $\boldsymbol{P}^{(c)}$ and enumerating each specific feature, the following loss can be obtained:

$$\mathcal{L}_{icl} = \sum_{v=1}^{V}\sum_{t=1}^{M}\sum_{t'=1}^{M} P_{t,t'}^{(v,c)} \ln\left(\frac{P_{tt'}^{(v,c)}}{\left(P_t^{(v)}\right)^{\alpha+1}\left(P_{t'}^{(c)}\right)^{\alpha+1}}\right), \tag{3}$$

where $P_t^{(v)}$ and $P_{t'}^{(c)}$ can be computed by summing over the t-th row and t'-th column of $\boldsymbol{P}^{(v,c)}$. In our experiments, we fix the balance parameter α to 9. Thus, the ultimate loss for extracting representations is $\mathcal{L}_1 = \mathcal{L}_{adv} + \mathcal{L}_{icl}$.

Consistent View-Specific Label Learning

Due to the heterogeneity of different views and their diverse importance for specific prediction tasks (Zhao et al. 2022), classifiers are independently established for each view representation to identify the most relevant multi-label subset. Classifiers F_v $(1 \leq v \leq V)$ are constructed for each specific representation, while F_0 is for the shared view. The confidence scores are given by $\boldsymbol{U}^v = F_v(\boldsymbol{h}_s^v)$ and $\boldsymbol{U}^0 = F_0(\boldsymbol{h}_c)$. Let $\widetilde{\boldsymbol{U}}^v$ $(0 \leq v \leq V)$ denote the predicted labels obtained with a 0.5 threshold. Multi-label samples inherently maintain an uneven label distribution, offering potential for meticulously learning view-specific labels based on label similarity (Yin and Zhang 2023). Moreover, the manifold assumption that similar samples should have similar labels (Wu et al. 2014) is crucial for improving classification, especially with diverse features and labels. Therefore, before integrating the view-specific results, we should fully consider the structural consistency between views and outcomes. Given that multi-label data is semantically represented using binary encoding and typically remains sparsity, the Jaccard distance (Park and Read 2019) is utilized to measure the similarity of two label sets from the perspective of their intersection and union. Denote the label similarity matrix as $\boldsymbol{T}^v$ $(0 \leq v \leq V)$ and its elements can be computed:

$$T_{i,j}^v = \frac{\langle \widetilde{\boldsymbol{U}}_i^v \cdot \widetilde{\boldsymbol{U}}_j^v \rangle}{\sum_{k=1}^{C} (\widetilde{U}_{i,k}^v + \widetilde{U}_{j,k}^v) - \langle \widetilde{\boldsymbol{U}}_i^v \cdot \widetilde{\boldsymbol{U}}_j^v \rangle}, \tag{4}$$

where $\langle \cdot \rangle$ denotes the vector dot product. Let $\boldsymbol{h}_c = \boldsymbol{h}_s^0$ and the similarity of two view features can be calculated in the cosine space (Yin and Sun 2022). To filter out unnecessary computations and enhance sensitivity in capturing similar samples, we leverage the principle in graph-based methods (Zhao et al. 2022) that considering similarity only within the neighborhood $N_p\left(\boldsymbol{h}_{s,i}^v\right)$. For sample i and j, their similarity in the extracted feature $\boldsymbol{h}_s^v$ $(0 \leq v \leq V)$ is defined as:

$$\boldsymbol{S}_{i,j}^v = \begin{cases} \left(\frac{\langle \boldsymbol{h}_{s,i}^v \cdot \boldsymbol{h}_{s,j}^v \rangle}{\|\boldsymbol{h}_{s,i}^v\| \|\boldsymbol{h}_{s,j}^v\|} + 1\right)/2, & \text{if } \boldsymbol{h}_{s,j}^v \in N_p\left(\boldsymbol{h}_{s,i}^v\right) \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

To ensure that similar sample pairs are close in the label space, we employ metric learning (Jia et al. 2020) for its strong discriminative capabilities. Let label similarity be the target and feature similarity be the learning object, the consistency loss can be formulated as:

$$\mathcal{L}_2 = \sum_{v=0}^{V} \sum_{i=1}^{N} \sum_{j \neq i}^{N} T_{i,j}^v S_{i,j}^v + (1 - T_{i,j}^v) \max(0, \text{Margin} - S_{i,j}^v), \tag{6}$$

where Margin is used to control the distance between non-matching sample pairs and we set it to the maximum 1.

Adaptively Label-Guided Transformer Fusion

In the real world, different view representations hold varying levels of importance for the same classification, and the same view contributes distinctly to each label in multi-label tasks (Zhuge et al. 2023). We select all views and six relatively balanced labels from Corel 5k and conduct experiments using the same classification method. As shown in Fig. 2, the contributions of different views to the same label and the same view across all labels are not uniform. In addition, multi-label correlations should be leveraged to explore the co-occurrence of label semantics and enhance model performance. To address these three considerations simultaneously during the result fusion, we employ a label-guided transformer for adaptive interaction awareness. Each category is mapped into the representation space and the self-attention mechanism is utilized to enable information interaction. Specifically, by concatenating the predictive confidence with each representation as auxiliary information, the imbalaced expressive power can be learned. To supplement the semantic information of label categories, C classtokens $\left\{\boldsymbol{cls}^i \in \mathbb{R}^{M+C}\right\}_{i=1}^{C}$ are randomly initialized before training and added to the input sequence. Therefore, the input sample tensor to the transformer is composed as follows:

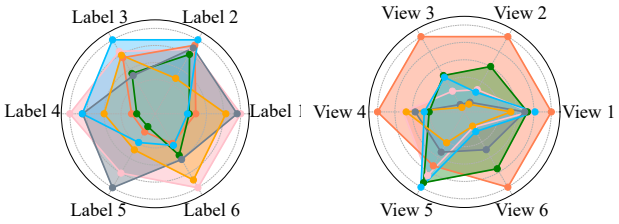


Figure 2: The results of different views under the same label and the same view for different labels on Corel 5k.

$$\hat{\boldsymbol{X}}_i = \left\{ \left[\boldsymbol{U}_i^0 \ \boldsymbol{h}_{c,i}\right], \cdots \left[\boldsymbol{U}_i^V \ \boldsymbol{h}_{s,i}^V\right], \boldsymbol{cls}_i^1, \cdots, \boldsymbol{cls}_i^C \right\}. \tag{7}$$

For each sample embedding, its queries, keys, and values are obtained using projective matrices $\left\{\boldsymbol{W}_t^q, \boldsymbol{W}_t^k, \boldsymbol{W}_t^v\right\}_{t=1}^{h}$ with h heads. The token-wise correlations $\boldsymbol{A}_t$ and output $\boldsymbol{B}_t$ are computed after information exchange:

$$\begin{aligned} \boldsymbol{A}_t &= \text{softmax}\left(\left(\hat{\boldsymbol{X}}_i \boldsymbol{W}_t^q\right)\left(\hat{\boldsymbol{X}}_i \boldsymbol{W}_t^k\right)^T / \sqrt{d_h}\right) \\ \boldsymbol{B}_t &= \boldsymbol{A}_t \left(\hat{\boldsymbol{X}}_i \boldsymbol{W}_t^v\right), \end{aligned} \tag{8}$$

where $d_h = (M + C)/h$. Fig. 3 illustrates the label-guided multi-head mechanism and it is worth noting that the feature-level, label-level and mapping-level interactions can be concurrently achieved by incorporating both predicted and actual label information into the self-attention process. On one hand, the importance of different view representations is adaptively adjusted, and individual views access structural association by linking with all subcategories, bringing them closer to the most relevant classtokens. On the other hand, cross-category information interaction implicitly promotes the learning of category correlations. After passing through the remaining transformer layers, the output is divided into two parts, i.e., the inferred results $\left\{\boldsymbol{U}^{0^*}, \boldsymbol{U}^{1^*}, \ldots \boldsymbol{U}^{V^*}\right\}$ from extracted representations and the predictions $\left\{\boldsymbol{cls}^{1^*}, \boldsymbol{cls}^{2^*}, \ldots, \boldsymbol{cls}^{C^*}\right\}$ generated by

each classtoken. Then we obtain the primary result by averaging the view-specific results $\boldsymbol{U}^m = \frac{1}{V+1}\sum_{i=0}^{V}\boldsymbol{U}^{i^*}$ and get the auxiliary prediction $\boldsymbol{U}^c$ by concatenating all referenced scores from classtokens. To promote the classification of multi-label, we adopt the following cross-entropy loss:

$$\mathcal{L}_{bce} = -\frac{1}{n_l C}\sum_{i\in\mathcal{F}}\sum_{j=1}^{C} Y_{i,j}\log\left(F_{i,j}\right) + \left(1-Y_{i,j}\right)\log\left(1-F_{i,j}\right), \tag{9}$$

where $F_{i,j}$ is the prediction. Thus, we can acquire the main classification loss $\mathcal{L}_{mc}$ and the ancillary loss $\mathcal{L}_{ac}$ according to the $\mathcal{L}_{bce}$ for $\boldsymbol{U}^m$ and $\boldsymbol{U}^c$. The final classification loss is $\mathcal{L}_3 = \mathcal{L}_{mc} + \mathcal{L}_{ac}$ and the total training loss is $\mathcal{L} = \mathcal{L}_1 + \lambda_1\mathcal{L}_2 + \lambda_2\mathcal{L}_3$, where λ_1 and λ_2 are penalty coefficients.

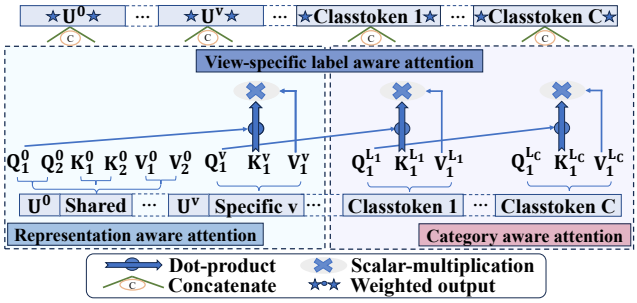


Figure 3: The label-guided self-attention with two heads.

Category-Enhanced Pseudo-Label Strategy

Since a large amount of unlabeled data exists, the model is prone to overfitting during training. Data augmentation is a powerful technique to improve the generalization capabilities of deep learning (Shorten and Khoshgoftaar 2019). Hence, the Mixup algorithm (Li, Li, and Yu 2024) is used to linearly interpolate pairs of training samples to generate new ones, which can reduce reliance on individual data points and provide a regularization effect. For each sample, the new training sample is produced with another random sample:

$$\begin{aligned}\boldsymbol{X}^v_{\text{new}} &= \alpha\boldsymbol{X}^v_{i\in\mathcal{F}} + (1-\alpha)\boldsymbol{X}^v_{j\in\mathcal{F},j\neq i}\\ \boldsymbol{Y}_{\text{new}} &= \text{sgn}\left(\alpha\boldsymbol{Y}_{i\in\mathcal{F}} + (1-\alpha)\boldsymbol{Y}_{j\in\mathcal{F},j\neq i} - 0.5\right),\end{aligned} \tag{10}$$

where $\alpha \sim \mathcal{U}(0,1)$ and $\text{sgn}(x) = 1$ for $x \geq 0$, otherwise, $\text{sgn}(x) = 0$. Owing to the capacity to supply new labeled data for training, pseudo-labeling methods have gained significant prominence in deep semi-supervised tasks (Jiang and Luo 2024). Due to the long-tail nature of multi-label distributions, the primary source of erroneous pseudo-labels is the excessive focus on the dominant labels (Meng et al. 2024). Therefore, the class-distribution-aware pseudo-label strategy is adopted to align the pseudo-labels closely with the true distribution and reduce noise. Considering each class $\boldsymbol{Y}_{:,k}$ separately in the unlabelled training samples, pseudo-labels can be obtained in the following way:

$$\hat{\boldsymbol{Y}}_{:,k} = \begin{cases} 1 & \text{if } \boldsymbol{U}^m_{:,k} \geq \tau(\alpha_k)\\ 0 & \text{if } \boldsymbol{U}^m_{:,k} \leq \tau(\beta_k)\\ -1 & \text{otherwise,}\end{cases} \tag{11}$$

where $\tau(\alpha_k)$ and $\tau(\beta_k)$ are two class-aware thresholds, and $\hat{Y}_{:,k} = -1$ indicates that the label will be excluded from training. Given that the class proportions of positive and negative labels in the labeled samples can tightly approximate the true class distribution (Xie et al. 2024), the solutions for determining $\tau(\alpha_k)$ and $\tau(\beta_k)$ can be addressed by

$$\begin{cases}\sum_{i\in\mathcal{R}}\mathbb{I}\left(U^m_{i,k} \geq \tau\left(\alpha_k\right)\right)/n_u = \hat{\gamma}_k\\ \sum_{i\in\mathcal{R}}\mathbb{I}\left(U^m_{i,k} \leq \tau\left(\beta_k\right)\right)/n_u = \hat{\rho}_k\end{cases} \tag{12}$$

where $\hat{\gamma}_k$ and $\hat{\rho}_k$ are the proportions of positive and negative labels in the labeled samples for class k. To further enhance the reliability of pseudo-labels, a feasible solution is to incorporate them after training the classifiers for some epochs and discard those with relatively low confidence. Therefore, top $\eta_1 \cdot \hat{\gamma}_k$ and $\eta_0 \cdot \hat{\rho}_k$ proportion probable pseudo-labels are selected for any class k, where η_1 and η_0 are used to control the confidence intervals and fixed at 0.8 in the experiments.

Experiment

Experimental Setting

Datasets and Evaluation metrics. Following (Liu et al. 2023b; Zhao et al. 2021), six popular multi-view multi-label datasets are selected in our experiments, i.e., Yeast (Guillaumin 2010), Corel 5k (Duygulu et al. 2002), VOC 2007 (Everingham et al. 2010), ESP Game (Ahn and Dabbish 2004), IAPR TC-12 (Grubinger et al. 2006), and MIR FLICKR (Huiskes and Lew 2008). Similar to (Liu et al. 2023a), four metrics commonly used in the multi-label learning are adopted as four evaluation criteria, i.e., Ranking Loss (RL), Accuracy (ACC), Average Precision (AP) and adapted area under curve (AUC). To facilitate performance comparison, we present 1-HL and 1- RL values in our report.

Comparison Methods. To validate the effectiveness of SMVTEP, we compare it with eight state-of-the-art approaches, i.e., iMVWL, TM3L, NAIML, DICNet, DD-IMvMLC, LMVCAT, ESMC and SCTML. The first six methods can address incomplete labels for MvMLC, while the last two can only handle the semi-supervised multi-label problem in a single view. Therefore, ESMC and SCTML are tested independently on each view with the best results reported. All parameters for comparison methods take precedence over the values recommended in their codes or papers.

Implementation Details. Each dataset is divided into training, validation and test sets in the ratio of 7:1:2. To simulate the semi-supervised situations, according to the pre-set labeled example ratio (LER), we randomly select LER% instances as labeled ones in the training set. All results are derived from ten independent runs and the final outcomes are presented as average values along with standard deviations. Our model is implemented by PyTorch on one NVIDIA GeForce RTX 4090 GPU of 24GB memory.

Experimental Results

Performance Evaluation. To comprehensively verify the capability of our SMVTEP in handling numerous unlabeled samples, we compare it against eight methods in

DATA	MET	SCTML	ESMC	iMVWL	TM3L	NAIM3L	DICNet	DIMC	LMVCAT	SMVTEP
YES	1-HL	76.26±1.04	57.56±1.19	69.10±2.87	69.77±0.41	64.15±0.25	69.76±0.23	69.54±2.57	75.17±0.89	**78.07±0.80**
	1-RL	78.02±0.86	73.01±1.26	68.93±6.15	72.18±2.93	64.02±0.29	75.76±0.95	68.54±4.69	76.98±1.19	**79.80±1.55**
	AP	70.65±1.44	63.95±1.94	59.90±5.49	65.05±2.74	56.74±0.60	67.51±0.55	61.46±4.82	69.23±1.10	**72.95±1.21**
	AUC	79.35±0.56	72.19±1.11	72.45±5.32	74.15±2.76	52.06±0.51	77.41±0.77	71.30±4.09	78.18±1.18	**80.70±1.50**
COR	1-HL	98.57±0.12	71.66±0.65	97.71±0.02	98.69±0.00	91.93±5.52	98.69±0.01	98.69±0.01	98.66±0.01	**98.70±0.00**
	1-RL	78.68±0.06	60.28±1.27	79.08±0.60	70.30±0.79	61.10±3.34	77.82±0.86	74.60±1.58	81.28±0.48	**84.80±0.92**
	AP	22.67±2.51	11.08±1.53	20.19±0.68	22.49±0.82	14.24±2.67	23.21±0.44	19.97±1.60	26.97±1.09	**32.52±0.54**
	AUC	79.93±0.25	60.27±1.35	79.31±0.64	77.91±0.27	61.50±3.24	78.01±0.82	74.63±1.78	81.45±0.52	**85.08±0.98**
VOC	1-HL	92.91±0.30	69.01±0.63	88.22±0.03	92.68±0.05	92.11±0.02	92.68±0.05	92.62±0.06	91.81±0.49	**93.17±0.02**
	1-RL	71.10±2.60	66.21±1.24	69.55±0.84	75.54±0.29	73.21±0.07	72.91±0.61	68.35±2.12	77.70±1.25	**82.04±1.57**
	AP	45.05±2.41	40.24±0.89	42.61±0.54	48.35±0.50	45.17±0.09	44.88±0.72	42.83±0.50	49.35±1.47	**54.18±1.03**
	AUC	74.69±2.38	66.89±1.45	71.95±0.36	80.37±0.36	62.93±0.32	75.65±0.65	71.95±1.47	79.77±1.26	**83.73±1.48**
ESP	1-HL	97.34±0.65	71.90±0.10	97.00±0.00	98.24±0.00	98.13±0.03	98.25±0.01	98.24±0.01	**98.26±0.01**	98.25±0.01
	1-RL	76.66±0.19	70.17±0.56	76.97±0.17	70.22±0.22	71.71±0.51	78.30±0.17	75.45±0.62	77.42±0.15	**81.23±0.50**
	AP	19.70±1.21	15.77±0.40	18.58±0.12	17.85±0.40	21.80±0.44	22.68±0.24	18.88±0.42	21.90±0.29	**27.22±0.63**
	AUC	77.10±0.27	70.07±0.53	77.54±0.19	72.93±0.31	71.80±0.44	78.50±0.13	75.83±0.63	77.88±0.11	**81.64±0.47**
IAP	1-HL	96.53±1.43	71.18±0.15	96.73±0.02	98.04±0.01	79.97±0.04	98.04±0.01	98.04±0.02	98.05±0.01	**98.06±0.01**
	1-RL	79.47±0.75	70.95±0.45	79.19±0.24	76.71±0.27	61.03±0.09	80.36±0.28	76.16±1.02	79.90±0.17	**84.31±0.71**
	AP	22.41±3.01	16.04±0.35	19.63±0.21	23.27±0.40	12.00±0.10	22.72±0.27	19.65±0.27	22.99±0.44	**27.76±0.21**
	AUC	80.01±0.42	70.80±0.40	79.55±0.21	79.07±0.34	54.83±0.07	80.29±0.18	76.33±0.97	80.28±0.25	**84.40±0.84**
MIR	1-HL	87.54±0.11	67.14±0.22	83.95±0.02	87.60±0.03	86.53+0.14	87.57±0.07	87.54±0.07	87.69±0.49	**89.19±0.11**
	1-RL	83.32±2.37	78.86±0.16	80.75±0.23	78.29±0.33	49.58+0.52	83.20±0.14	79.08±1.27	85.15±0.56	**87.27±0.21**
	AP	51.99±7.19	42.81±0.41	45.15±0.75	50.01±0.55	45.30+1.92	51.69±0.37	45.18±1.18	58.19±1.10	**60.27±0.25**
	AUC	82.94±2.47	76.94±0.24	80.45±0.12	81.11±0.26	58.56+0.24	81.71±0.17	78.01±1.41	83.86±0.37	**85.53±0.28**

Table 1: 1-HL, 1-RL, AP and AUC of different methods on six public datasets with LER fixed to 6%. The best result on each row is bolded and the second-best result is underlined.

scenarios where unlabeled samples greatly outnumber labeled ones. Besides, for label insufficient, the value of LER is fully considered, being taken at different ratios of $\{3\%, 6\%, 9\%, 12\%, 15\%, 18\%, 21\%\}$. Tabel 1 displays the four metrics with LER fixed at 6%, while Fig. 4 illustrates the variation of AP when LER changes.

Regarding the comparison results, we have the following observations: 1) Our method achieves better among all compared methods in almost all cases, particularly excelling in the most representative AP metric. 2) Compared to the semi-supervised methods ESMC and SCTML, our SMVTEP reveals substantial advantages, such as over 50% improvement in AP on Corel 5k and VOC 2007. Additionally, our approach also outperforms methods designed for incomplete MvMLC. For instance, Table 1 illustrates that our SMVTEP reliably enhances performance even against LMVCAT, the top-performing MvMLC method on MIR FLICKR. Overall, our method exhibits a strong capability to simultaneously address the challenges of numerous features and label categories, as well as insufficient annotations. 3) With LER increasing from 3% to 21%, SMVTEP still outperforms the other eight methods. Besides, our method exhibits robust performance and consistently achieves relatively promising results with lower LER. For example, SMVTEP and the second-best method LMVCAT achieve AP of 29.72% and 27.43% when LER=21% on ESP Game. As LER=3%, the performance of SMVTEP is 25.05% and superior to 20.05% of LMVCAT, which indicates that our method is better suitable for situations with a high lack of annotations.

Ablation Study. The ablation experiments are conducted to thoroughly examine the influence of the critical modules of SMVTEP. Specifically, we evaluate the impact of removing Generative Adversarial Networks (S_1), the label-guided transformer (S_2), and the pseudo-label strategy (S_3) by using raw views, averaging view-specific predictions, and directly assigning classifier results as pseudo-labels, respectively. As can be seen in Table 2, the methods of feature semantic extraction and late fusion are beneficial for enhancing performance. Moreover, our pseudo-label strategy plays a crucial role, especially with larger datasets. While noise accumulation in pseudo-labels severely hinders classification on Corel 5k and VOC 2007, our method effectively improves pseudo-label quality by aligning distributions.

S_1	S_2	S_3	Yeast		Corel 5k		VOC 2007	
			AP	AUC	AP	AUC	AP	AUC
✗	✓	✓	71.93	81.01	31.34	83.49	51.65	81.99
✓	✗	✓	72.05	80.91	29.97	79.45	50.77	77.02
✓	✓	✗	71.11	80.46	20.59	77.32	43.04	72.54
✓	✓	✓	**75.51**	**82.49**	**32.09**	**84.47**	**54.41**	**84.49**

Table 2: Ablation study on Yeast, Corel 5k and VOC 2007 with LER=6%. '✓' and '✗' represent the used and not used corresponding item, respectively.

Parameter Sensitivity and Convergence. To study the impact of λ_1 and λ_2, we present the performance under dif-

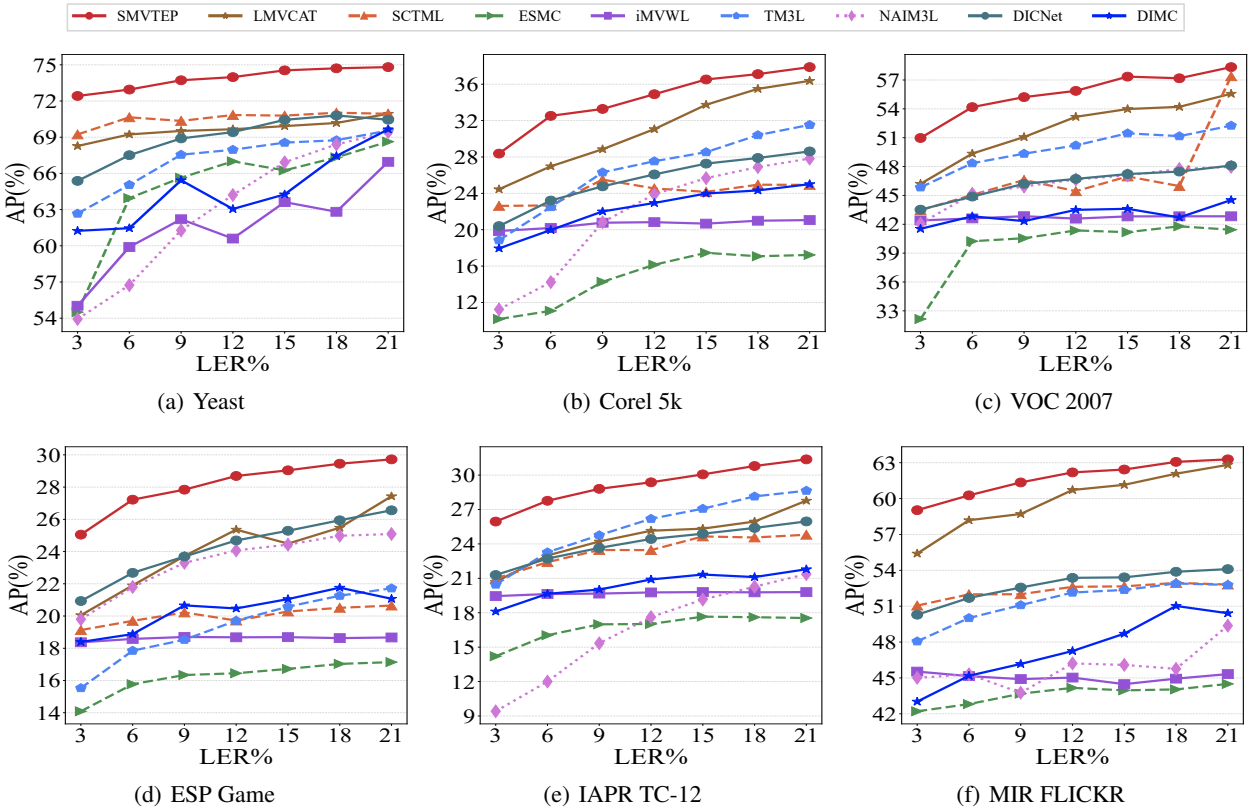
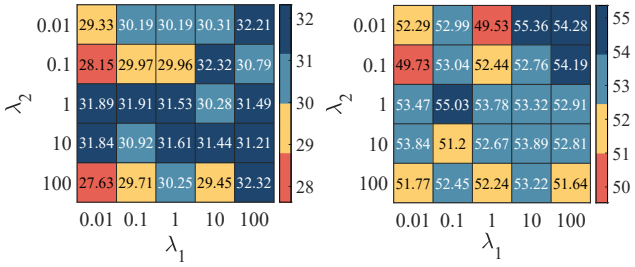


Figure 4: AP comparisons on six datasets with LER varying from 3% to 21%.

ferent parameter combinations in Fig. 6. The heatmap reveals that the performance improves as λ_1 gets closer to 100 and λ_2 approaches 0.01 and our SMVTEP is not so sensitive to both parameters. We also report the variation trends of both the training loss and AP value and find that their convergence is nearly concurrent. Besides, the performance markedly improves after incorporating pseudo-labels.



(a) Corel 5k (b) VOC 2007

Figure 5: Parameter analysis of the trade-off parameters λ_1 and λ_2 on Corel 5k and VOC 2007.

Conclusion

To tackle the problem of numerous unlabeled samples under diverse features and labels, we propose a novel deep semi-supervised learning method named SMVTEP in this paper.

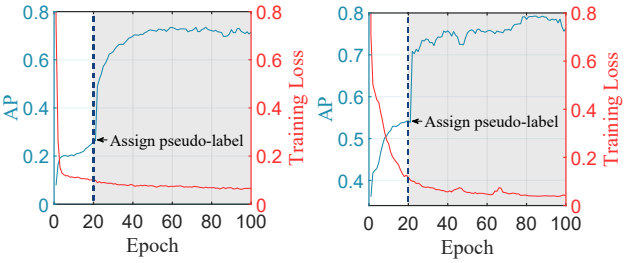


Figure 6: The convergence of our SMVTEP during training on Corel 5k and VOC 2007.

SMVTEP leverages Generative Adversarial Networks to extract semantically distinct representations, performs view-specific label learning with manifold constraints, and employs a label-guided transformer for late fusion to simultaneously explore internal relationships within feature and label spaces and mapping semantics. Moreover, SMVTEP adopts a combination of data augmentation and a category-enhanced pseudo-label strategy to handle sparse annotations and extract high-quality supervisory information. Finally, extensive experimental results demonstrate the superiority of SMVTEP. In the future, we will extend to address the problems of label noise and class imbalance.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62376281, the NSF for Distinguished Young Scholars under Grant No. 62425607, and the Key NSF of China under Grant No. 62136005.

References

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

Akbarnejad, A. H.; and Baghshah, M. S. 2018. An efficient semi-supervised multi-label classifier capable of handling missing labels. *IEEE Transactions on Knowledge and Data Engineering*, 31(2): 229–242.

Chang, W.-C.; Yu, H.-F.; Zhong, K.; Yang, Y.; and Dhillon, I. S. 2020. Taming pretrained transformers for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD International Conference on knowledge Discovery and Data Mining*, 3163–3171.

Dong, H.-C.; Li, Y.-F.; and Zhou, Z.-H. 2018. Learning from semi-supervised weak-label data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2926–2933.

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

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88: 303–338.

Grubinger, M.; Clough, P.; Müller, H.; and Deselaers, T. 2006. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop onto Image*, volume 2, 1–11.

Guan, R.; Li, Z.; Tu, W.; Wang, J.; Liu, Y.; Li, X.; Tang, C.; and Feng, R. 2024. Contrastive Multiview Subspace Clustering of Hyperspectral Images Based on Graph Convolutional Networks. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–14.

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

Han, X.; Zhou, F.; Ren, Z.; Wang, X.; and You, X. 2024. View-specific anchors coupled tensorial bipartite graph learning for incomplete multi-view clustering. *Information Sciences*, 664: 120335.

Hu, D.; Dong, Z.; Liang, K.; Yu, H.; Wang, S.; and Liu, X. 2024a. High-order Topology for Deep Single-cell Multi-view Fuzzy Clustering. *IEEE Transactions on Fuzzy Systems*.

Hu, D.; Liu, S.; Wang, J.; Zhang, J.; Wang, S.; Hu, X.; Zhu, X.; Tang, C.; and Liu, X. 2024b. Reliable Attribute-missing Multi-view Clustering with Instance-level and feature-level Cooperative Imputation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1456–1466.

Hu, P.; Zhen, L.; Peng, X.; Zhu, H.; Lin, J.; Wang, X.; and Peng, D. 2023. Deep supervised multi-view learning with graph priors. *IEEE Transactions on Image Processing*, 33: 123–133.

Hu, S.; and Miao, D. 2022. Multi granularity based label propagation with active learning for semi-supervised classification. *Expert Systems with Applications*, 192: 116276.

Huiskes, M. J.; and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, 39–43.

Jia, X.; Jing, X.-Y.; Zhu, X.; Chen, S.; Du, B.; Cai, Z.; He, Z.; and Yue, D. 2020. Semi-supervised multi-view deep discriminant representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7): 2496–2509.

Jiang, Z.; and Luo, T. 2024. Deep Incomplete Multi-View Learning Network with Insufficient Label Information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 12919–12927.

Jung, S.; Jeoung, J.; Hong, T.; and Jang, H. 2024. Vision-based multi-label detection framework for capturing occupant action and clothing information using large-scale dataset. *Building and Environment*, 257: 111537.

Li, J.; Li, G.; and Yu, Y. 2024. Inter-domain mixup for semi-supervised domain adaptation. *Pattern Recognition*, 146: 110023.

Li, J.; Zhu, X.; Wang, H.; Zhang, Y.; and Wang, J. 2024. Stacked co-training for semi-supervised multi-label learning. *Information Sciences*, 677.

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.

Lin, Y.; Gou, Y.; Liu, X.; Bai, J.; Lv, J.; and Peng, X. 2022. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4447–4461.

Liu, C.; Wen, J.; Luo, X.; Huang, C.; Wu, Z.; and Xu, Y. 2023a. Dicnet: Deep instance-level contrastive network for double incomplete multi-view multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8807–8815.

Liu, C.; Wen, J.; Luo, X.; and Xu, Y. 2023b. Incomplete multi-view multi-label learning via label-guided masked view-and category-aware transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8816–8824.

Meng, L.; Zhang, Q.; Yang, R.; and Huang, Y. 2024. Unsupervised Deep Hashing with Dynamic Pseudo-Multi-Labels for Image Retrieval. *IEEE Signal Processing Letters*, 31: 909–913.

Miller, H. A.; and Lowengrub, J. 2022. Modeling of tumor growth with input from patient-specific metabolomic data. *Annals of biomedical engineering*, 50(3): 314–329.

Ou, S.; Xue, Z.; Li, Y.; Liang, M.; Cai, Y.; and Wu, J. 2024. View-Category Interactive Sharing Transformer for Incomplete Multi-View Multi-Label Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27467–27476.

Park, L. A.; and Read, J. 2019. A blended metric for multi-label optimisation and evaluation. In *Machine Learning and Knowledge Discovery in Databases: European Conference*, 719–734. Springer.

Peng, X.; Zhu, H.; Feng, J.; Shen, C.; Zhang, H.; and Zhou, J. T. 2019. Deep clustering with sample-assignment invariance prior. *IEEE Transactions on Neural Networks and Learning Systems*, 31(11): 4857–4868.

Shorten, C.; and Khoshgoftaar, T. M. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1): 1–48.

Su, X.; and Xu, Y. 2024. Imbalanced and missing multi-label data learning with global and local structure. *Information Sciences*, 120910.

Sun, F.; and Xie, M.-K. 2024. A Deep Model for Partial Multi-label Image Classification with Curriculum-based Disambiguation. *Machine Intelligence Research*, 1–14.

Sun, N.; Luo, T.; Zhuge, W.; Tao, H.; Hou, C.; and Hu, D. 2021. Semi-supervised learning with label proportion. *IEEE Transactions on Knowledge and Data Engineering*, 35(1): 877–890.

Sun, S.; and Wang, B. 2024. Decoupled representation for multi-view learning. *Pattern Recognition*, 151: 110377.

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

Wen, J.; Liu, C.; Deng, S.; Liu, Y.; Fei, L.; Yan, K.; and Xu, Y. 2023. Deep double incomplete multi-view multi-label learning with incomplete labels and missing views. *IEEE Transactions on Neural Networks and Learning Systems*, 35: 11396–11408.

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

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 *International Joint Conference on Artificial Intelligence*, 3884–3890.

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

Xie, M.-K.; Xiao, J.; Liu, H.-Z.; Niu, G.; Sugiyama, M.; and Huang, S.-J. 2024. Class-distribution-aware pseudo-labeling for semi-supervised multi-label learning. *Advances in Neural Information Processing Systems*, 36.

Xin, B.; Lu, S.; Wang, Q.; Deng, F.; Shi, X.; Cheng, J.; and Kang, Y. 2023. Simultaneous scheduling of processing machines and automated guided vehicles via a multi-view modeling-based hybrid algorithm. *IEEE Transactions on Automation Science and Engineering*, 21: 4753–4767.

Yin, J.; and Sun, S. 2022. Incomplete multi-view clustering with cosine similarity. *Pattern Recognition*, 123: 108371.

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

Zhan, W.; and Zhang, M.-L. 2017. Inductive semi-supervised multi-label learning with co-training. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 1305–1314.

Zhang, J.; Li, S.; Jiang, M.; and Tan, K. C. 2020. Learning from weakly labeled data based on manifold regularized sparse model. *IEEE Transactions on Cybernetics*, 52(5): 3841–3854.

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

Zhao, D.; Gao, Q.; Lu, Y.; and Sun, D. 2021. Two-step multi-view and multi-label learning with missing label via subspace learning. *Applied Soft Computing*, 102: 107120.

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

Zhu, C.; Liu, Y.; Miao, D.; Dong, Y.; and Pedrycz, W. 2023. Within-cross-consensus-view representation-based multi-view multi-label learning with incomplete data. *Neurocomputing*, 557: 126729.

Zhuge, W.; Luo, T.; Fan, R.; Tao, H.; Hou, C.; and Yi, D. 2023. Absent multiview semisupervised classification. *IEEE Transactions on Cybernetics*, 54(3): 1708–1721.