Quality-aware and Soft Consistency Driven Representation Fusion for Incomplete Multi-view Multi-label Classification

Yadong Liu[1], Waikeung Wong[2], Yulong Chen[1], Jie Wen[1]*

[1]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China
[2]School of Fashion and Textiles, Hong Kong Polytechnic University
liuyadong221010@163.com, calvinwong@aidlab.hk, chenyulonghit@163.com, jiewen_pr@126.com,

Abstract

Multi-view multi-label classification aims to utilize the rich information contained in multiple views for accurate classification. However, in real-world applications, its performance is often severely constrained by the concurrent missingness of both views and labels. To address this problem, this paper first targets the drawback of representation degradation in traditional feature disentanglement methods caused by strong consistency constraints and proposes a soft consistency constraint. This constraint not only effectively aligns the shared information and maximally avoids the compression of information beneficial to the classification task, but it also enhances the aggregation effect of high-quality representations on other representations. Furthermore, to address the coarse-grained problem of traditional fusion strategies, we designed a quality assessment network that achieves instance-level dynamic weighted fusion in a data-driven manner. Extensive experiments on multiple benchmark datasets demonstrate that our method achieves state-of-the-art performance in both incomplete and complete data scenarios, showcasing its robustness and generality.

Introduction

Rapid advancement of information technology has led to real-world data becoming increasingly available in multi-modal and multi-view formats (Lu et al. 2024). For example, a patient's medical record might include both CT scans and X-ray images. This heterogeneous data, drawn from diverse sources, provides complementary information and varied perspectives to understand complex phenomena. Consequently, multi-view learning has found widespread application in various domains, such as medical diagnosis (Lin et al. 2025; Luo et al. 2025), cross-view retrieval (Wang et al. 2020, 2023b; Su et al. 2025), and unsupervised tasks like multi-view clustering (Li et al. 2024, 2025).

However, a prevalent and pressing challenge in applying multi-view learning theories to the real-world is data incompleteness (Zhang et al. 2024). Due to a variety of practical factors, such as data acquisition failures, storage limitations, or privacy concerns, the samples we obtain often suffer from missing data in one or more views. This issue causes traditional multi-view methods, which rely on the assumption of data completeness, to suffer a sharp degradation in performance, or even fail entirely, when processing incomplete data. Consequently, their practical utility is severely compromised. Therefore, increasing attention is being directed towards how to handle the more prevalent and challenging real-world problem of missing multi-view data. The research direction dedicated to solving such issues is collectively known as Incomplete Multi-view Learning (IMVL).

Multi-label classification (Zhang and Zhou 2013) is more general-purpose than single-label classification, as it accurately models real-world scenarios where an object can possess multiple attributes simultaneously. However, this greater applicability introduces unique challenges. For example, a social media post may only be annotated by users with a subset of its relevant labels. Consequently, learning effectively from this incomplete supervisory information has become one of the challenges in the field. The objective of Incomplete Multi-view Missing Multi-label Classification (iM3C) is to train a model under the dual challenge of missing views and labels and to be able to achieve accurate and robust multi-label prediction for samples in a testing phase where some of views are still missing. As observed in several previous studies (Liu et al. 2024a; Xie et al. 2024), Although both missing views and missing labels are challenges in incomplete learning, their negative impact on model performance is not equivalent. At the same missing rate, the absence of view information typically leads to a performance degradation more significant than the absence of labels. In light of this, the research focus of this paper is on how to maximally extract and utilize the latent, task-relevant semantic information from incomplete multi-view data, in order to address the severe challenges posed by missing views.

In recent years, a dominant research direction in this field has been learning a consistent representation that is shared across all views (Lv et al. 2022; Wang et al. 2023a; Liu et al. 2024c). The core idea is that consistent information is redundantly encoded across multiple views. Consequently, even when some views are missing, the model can still effectively reconstruct or infer this shared representation from the available views to perform downstream tasks. However, the effectiveness of such strategies, which are based solely on shared information, is based on a strong assumption (Tsai et al. 2020) that: shared information across views is sufficient to fulfill the requirements of the downstream task. In

*Corresponding author.

many real-world scenarios, this assumption is often overly optimistic. Each view not only contains shared information, but also carries a substantial amount of view-specific private information. Moreover, this private information can contain discriminatory features that are vital for the downstream task (Liu et al. 2023a). Our research reveals a trade-off in multi-view representation learning: the inherent tension between consistency and informativeness. When a model excessively pursues consistency among view representations, it may do so at the cost of informativeness. This can lead the model to actively suppress or discard view-specific information that is useful for the downstream task, ultimately resulting in a degradation of representation quality.

Therefore, we introduce a soft consistency regularization. In contrast to hard constraints that mandate precise alignment, our approach fosters consistency among shared inter-view information while concurrently safeguarding the existence of task-relevant, view-specific private information, enhances the aggregation effect of high-quality representations on other representations. After obtaining the representations for each view, the challenge lies in how to fuse them effectively. An ideal fusion strategy should aim to strike an optimal balance between retaining task-relevant information and compressing redundant information. To address the coarse-grained and non-adaptive problems prevalent in existing fusion strategies, we propose a quality-aware, instance-level dynamic fusion scheme. Unlike conventional methods that treat all samples equally, our scheme is capable of generating dynamic fusion weights for each sample based on its unique information, thereby achieving a more fine-grained and efficient information integration. We name our **QU**ality-**A**ware representation **L**earning v**I**a sof**T** consistenc**Y** based method as **QUALITY**. In summary, the main contributions of this paper can be summarized as follows:

- We propose a soft consistency constraint that not only aligns shared information across views but also preserves task-relevant private information from over-compression.
- We design a quality-aware fusion module that moves beyond static, coarse-grained approaches by generating a dynamic, adaptive fusion strategy tailored to each sample's unique quality.
- Extensive experiment results on multiple datasets show that our proposed method outperforms current state-of-the-art methods, not only in various incomplete data scenarios but also in settings with complete data.

Problem Definition and Related Works

Notation

To better describe the problem and our proposed method, we define the raw data set as $\mathbf{X} = \{\mathbf{x}_i^{(v)} \mid i = 1, ..., \mathbf{n}; v = 1, ..., \mathbf{m}\}$ with $\mathbf{m}$ views and $\mathbf{n}$ samples, where $\mathbf{x}_i^{(v)} \in \mathbb{R}^{d_v}$ is the v-th view of i-th sample, and d_v denotes the dimensionality. Consequently, we define $\mathbf{Y} \in \{0,1\}^{\mathbf{n}\times\mathbf{c}}$ the label set, where $\mathbf{c}$ is the number of categories. Under the challenging setting of simultaneous view and label missingness, we define $\{\mathcal{V}_i\}_{i=1}^{\mathbf{n}}$ as the known view set, and we also define $\{\mathcal{G}_i\}_{i=1}^{\mathbf{n}}$ as the known label set. Specially, $\mathcal{V}_i$ denotes the views that can be observed of i-th sample, and $\mathcal{G}_i$ denotes the labels that can be observed of i-th sample.

Incomplete Multi-view Learning

Current approaches to Incomplete Multi-view Learning (IMVL) are primarily twofold. One line of work focuses on learning representations via observed views. This approach aims to effectively distill and synthesize salient discriminative features from the observed views for the purpose of enhancing performance on downstream tasks. The other line of work centers on missing view imputation, which seeks to first explicitly generate the missing data using the available views and then perform learning on the completed dataset. To balance interview consistency and diversity, the model named CDMM (Zhao et al. 2021) uses view-specific classifiers to align predictions while leveraging the Hilbert-Schmidt Independence Criterion (HSIC) to retain view-specific information. Based on the structural consistency between view and label spaces, the model named LVSL (Zhao et al. 2022) simultaneously models multi-view consistency and low-rank inter-label correlations. Subsequently, at the prediction stage, it leverages view-specific weights to capture complementary information. Within the domain of generative methods, the LaSA model (Liu et al. 2024b) leverages the inherent graph structure of the samples to guide the reconstruction of missing view data.

Multi-label Learning with Missing Labels

In a manner highly analogous to Incomplete Multi-view Learning, the prevailing methods for Multi-label Learning with Missing Labels can be classified into two major paradigms. The first, association mining from known labels, seeks to uncover the latent correlations within the set of observed labels and then utilize these relationships to infer the state of the full label space. The second paradigm, generative-based missing label imputation, is more direct. It actively attempts to reconstruct the absent labels first, after which the model is trained on the fully imputed label information. Taking into account the complementary nature of global and local information, the DM2L (Ma and Chen 2021) model captures both the global low-rank label structure and the local low-rank label structure within the original label space. This approach provides enhanced label discriminability. By learning a unique, dynamic threshold for each class, the CAP (Xie et al. 2023) model effectively overcomes the limitations of conventional pseudo-labeling strategies when dealing with multi-label correlations and an unknown quantity of labels.

iM3C

Although there are numerous effective methods for handling missing views or missing labels individually, they are difficult to apply directly to the iM3C problem, which involves dual incompleteness of both views and labels. To address the iM3C problem, the iMVWL model (Tan et al. 2018) co-trains three main components within a unified framework: a shared subspace for view alignment, a local structure for mining label correlations, and a final classifier. The strategy

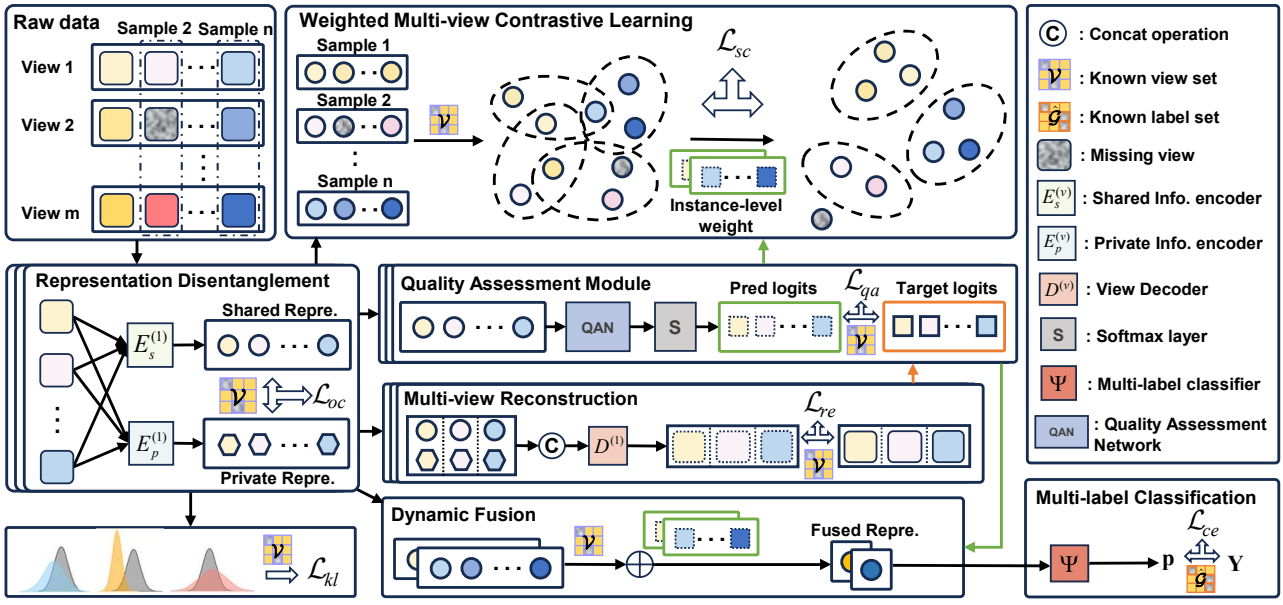


Figure 1: The main framework of our QUALITY. The innovation of our framework comprises two synergistic modules: a quality assessment module and a weighted soft consistency constraint module. Among them, the quality assessment module generates a dynamic weight for each view representation. This weight subsequently plays a dual role: it is used not only to guide the instance-level weighted fusion but also to modulate the contrastive loss function for the consistency constraint.

of the NAIM3L model (Li and Chen 2021) is to learn a consistent multi-view representation to counter view missingness, while simultaneously leveraging both global and local inter-label structures to compensate for the lack of supervisory information. DICNet (Liu et al. 2023b) successfully applies contrastive learning to the iM3C problem, demonstrating the great potential of this technical direction by learning unified, robust representations from doubly incomplete data.

Method

In this section, we will detail our proposed method, which consists of four components: multi-view representation disentanglement, shared soft consistency constraint, dynamic instance-level fusion, and multi-label classification.

Multi-view Representation Disentanglement

Although multi-view data contains richer information, traditional fusion methods often suffer performance bottlenecks due to their inability to effectively handle significant information redundancy (Hwang et al. 2021). Some works (Liu et al. 2023a; Chen et al. 2023) propose to solve the information redundancy through feature disentanglement, which decomposes the features of each view into two parts: view-shared and view-private components. In the decision-making phase, the model primarily relies on the more robust shared features and combines them with the private features to make the final judgment. Based on this, we designed a multi-view disentanglement network. Taking the v-th view of i-th sample as an example, the raw data are processed by two distinct encoders to separately extract the shared and private information, i.e., $E_s^{(v)}(\mathbf{x}_i^{(v)}) = (\mu_{i,s}^{(v)}, \sigma_{i,s}^{(v)})$ and $E_p^{(v)}(\mathbf{x}_i^{(v)}) = (\mu_{i,p}^{(v)}, \sigma_{i,p}^{(v)})$, where $E_s^{(v)}$ and $E_p^{(v)}$ denote the shared information encoder and the view private information encoder. The shared representation $\mathbf{z}_{i,s}^{(v)}$ and the private representation $\mathbf{z}_{i,p}^{(v)}$ can be sampled from the Gaussian distributions $\mathcal{N}(\mu_{i,s}^{(v)}, (\sigma_{i,s}^{(v)})^2)$ and $\mathcal{N}(\mu_{i,p}^{(v)}, (\sigma_{i,p}^{(v)})^2)$, respectively. To ensure information integrity during the disentanglement process, we concatenate the shared and private representations and feed them into a decoder to reconstruct the raw data. The loss function of this reconstruction task acts as a self-supervised constraint, ensuring that the two disentangled components collectively preserve the entirety of the original information:

$$\mathcal{L}_{re} = \frac{1}{\sum_{i=1}^{\mathbf{n}} |\mathcal{V}_i|} \sum_{i=1}^{\mathbf{n}} \sum_{v \in \mathcal{V}_i} ||\mathbf{x}_i^{(v)} - \hat{\mathbf{x}}_i^{(v)}||_2^2, \tag{1}$$

where $\hat{\mathbf{x}}_i^{(v)} = D^{(v)}(\text{concat}(\mathbf{z}_{i,s}^{(v)}, \mathbf{z}_{i,p}^{(v)}))$ and $D^{(v)}$ is the decoder of the v-th view. Additionally, we regularize the posterior distributions of the shared and private representations, respectively, by calculating the KL divergence loss between them and a standard normal distribution:

$$\mathcal{L}_{kl} = \frac{1}{\sum_{i=1}^{\mathbf{n}} |\mathcal{V}_i|} \sum_{i=1}^{n} \sum_{v \in \mathcal{V}_i} [D_{KL}\left(q_s^{(v)}(\mathbf{z}_{i,s}^{(v)}|\mathbf{x}_i^{(v)})||p(\mathbf{z}_{i,s}^{(v)})\right) + D_{KL}\left(q_p^{(v)}(\mathbf{z}_{i,p}^{(v)}|\mathbf{x}_i^{(v)})||p(\mathbf{z}_{i,p}^{(v)})\right)], \tag{2}$$

where $q_s^{(v)}(\mathbf{z}_{i,s}^{(v)}|\mathbf{x}_i^{(v)})$ and $q_p^{(v)}(\mathbf{z}_{i,p}^{(v)}|\mathbf{x}_i^{(v)})$ are predicted by two encoders. Furthermore, to ensure information disentanglement between the shared and private representations, we

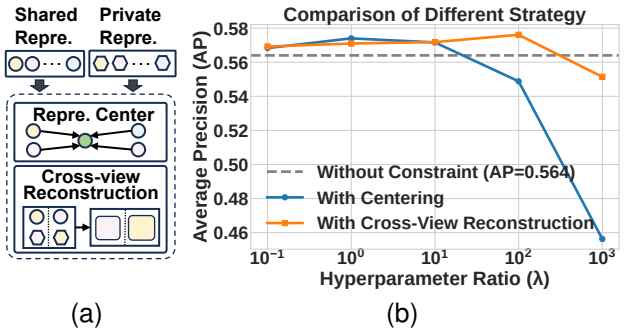


Figure 2: (a) Two Consistency Constraint Methods. (b) Performance of the Two Constraints with Varying Hyperparameters on Pascal07 dataset.

impose an orthogonality constraint. The objective of this constraint is to minimize the linear correlation between the two representations, thereby ensuring that they learn information from dimensions that are independent of each other:

$$\mathcal{L}_{oc} = \frac{1}{\sum_{i=1}^{\mathbf{n}} |\mathcal{V}_i|} \sum_{i=1}^{\mathbf{n}} \sum_{v \in \mathcal{V}_i} ((\mathbf{z}_{i,s}^{(v)})^T \mathbf{z}_{i,p}^{(v)})^2. \tag{3}$$

where $\mathbf{z}_{i,s}^{(v)}$ and $\mathbf{z}_{i,p}^{(v)}$ are representations after normalization. The multi-view disentanglement loss is as follows:

$$\mathcal{L}_{dis} = \alpha \mathcal{L}_{re} + \gamma \mathcal{L}_{kl} + \mathcal{L}_{oc} \tag{4}$$

where α and γ are trade-off parameters.

Shared Soft Consistency Constraint

Although multi-view data describe the same object, the inherent heterogeneity of each view in terms of feature space, data distribution, and noise characteristics makes their direct fusion extremely difficult (Zhang et al. 2024). Some works (Zeng et al. 2023; Liu et al. 2024c) are based on the shared information assumption, which posits that the shared multi-view representation is sufficient for the needs of downstream tasks. Therefore, after completing feature disentanglement, these methods use the shared representation to perform the downstream task. Based on this, we explored two consistency constraint schemes to extract shared multi-view representations, as shown in Fig. 2(a): the first is by applying a centering constraint to the representations, and the second is by constraining representation consistency through cross-view reconstruction. As shown in Fig. 2(b), we investigated the weight of the consistency constraint. The hyperparameter λ is a trade-off coefficient that balances the consistency constraint and the reconstruction loss. The results indicate that its contribution to performance improvement is limited; moreover, when the constraint is too strong, it leads to performance degradation.

Therefore, inspired by (Wu et al. 2024), we employ contrastive learning to impose a soft consistency constraint. It only pulls the representations of multiple views from the same sample closer, rather than forcing them to be identical. This allows for the alignment of shared information while also permitting the shared representation to contain some non-shared information that is beneficial for the downstream task. To account for the varying quality of views across different samples, we incorporate a dynamic weight $\hat{\mathbf{q}}$ at the instance-level, into our contrastive loss function. The purpose of this weight is to boost the aggregating influence of high-quality representations over others. The details of how $\hat{\mathbf{q}}$ is generated will be detailed in the following section. For any pair of views (k, l), the similarity between the i-th sample of view k and the j-th sample of view l is calculated by the following formula:

$$\mathbf{s}_{i,j}^{(k,l)} = \frac{((\mathbf{z}_{i,s}^{(k)})^T \mathbf{z}_{j,s}^{(l)}) \cdot (\hat{\mathbf{q}}_{i,s}^{(k)} \cdot \hat{\mathbf{q}}_{j,s}^{(l)})}{\tau}, \tag{5}$$

where τ is the temperature. For the i-th sample of view k, the InfoNCE loss is calculated as follows:

$$\mathcal{L}_i^{(k,l)} = -\mathbf{s}_{i,i}^{(k,l)} + \log\Big(\sum_{j \in \mathcal{V}^{k,l}} \exp(\mathbf{s}_{i,j}^{(k,l)})\Big). \tag{6}$$

The final soft consistency constraint loss is obtained by calculating and then averaging the weighted contrastive losses over all possible view pairs, with each loss computed on its respective subset of valid samples:

$$\mathcal{L}_{sc} = \frac{\sum_{k=1}^{\mathbf{m}} \sum_{l=k+1}^{\mathbf{m}} \sum_{i \in \mathcal{V}^{k,l}} \mathcal{L}_i^{(k,l)}}{\sum_{k=1}^{\mathbf{m}} \sum_{l=k+1}^{\mathbf{m}} |\mathcal{V}^{k,l}|}, \tag{7}$$

where $\mathcal{V}^{k,l}$ denotes the set of samples for which both view k and view l are available.

Dynamic Instance-Level Fusion

Traditional fusion methods, such as mean fusion or view-level weighting, have a fundamental limitation: they implicitly assume that the importance of each view is fixed for all samples. However, in real-world scenarios, the quality and relevance of the views are dynamic and variable among samples. Therefore, such a static fusion strategy struggles to adapt to the true data distribution, thereby limiting the model's upper performance bound. Another line of work (Liu et al. 2025) assesses the quality of view representations indirectly, which is achieved by training a group of separate single-view classifiers along with a quality discriminator network. Inspired by this, we have designed a novel dynamic instance-level weighted fusion strategy. Unlike the static and coarse-grained fusion of traditional methods, this strategy can adaptively generate a unique set of fusion weights for each sample based on the quality of its view representations. Taking the i-th sample as an example, for the shared representation of each view, we calculate its average log variance to serve as a measure of its uncertainty $\{\mathbf{u}_{i,s}^{(v)}\}_{v \in \mathcal{V}}$. Then, the shared representation is concatenated with the uncertainty metric to obtain an augmented feature vector:

$$\hat{\mathbf{z}}_{i,s}^{(v)} = \text{concat}(\mathbf{z}_{i,s}^{(v)}, \mathbf{u}_{i,s}^{(v)}). \tag{8}$$

Next, the feature vector is fed into a quality assessment network f_θ. The output layer of this network is a Sigmoid function, which maps the feature vector to a quality score:

$\mathbf{q}_{i,s}^{(v)} = f_\theta(\hat{\mathbf{z}}_{i,s}^{(v)})$. The normalized quality score for its v-th view is given by:

$$\hat{\mathbf{q}}_{i,s}^{(v)} = \frac{\mathbf{q}_{i,s}^{(v)}}{\sum_{v' \in \mathcal{V}_i} \mathbf{q}_{i,s}^{(v')} + \epsilon}, \tag{9}$$

where $\epsilon > 0$ is a small value added for numerical stability. Subsequently, we utilize the generated quality scores as weights to perform a weighted fusion of the means and variances from all view representations, thereby obtaining a unified, quality-aware fused representation.

$$\hat{\mu}_{i,s} = \sum_{v \in \mathcal{V}_i} \hat{\mathbf{q}}_{i,s}^{(v)} \cdot \mu_{i,s}^{(v)}, \quad (\hat{\sigma}_{i,s})^2 = \sum_{v \in \mathcal{V}_i} (\hat{\mathbf{q}}_{i,s}^{(v)})^2 \cdot (\sigma_{i,s}^{(v)})^2. \tag{10}$$

We designed a training strategy for the quality assessment network, the core of which is to use the reconstructability of a feature as a proxy for its quality. Specifically, we impose a constraint by minimizing the difference between the normalized quality score predicted by the network $\hat{\mathbf{q}}_{i,s}^{(v)}$ and a target score derived from the reconstruction loss. This target score is defined as the negative exponential function of the sample's normalized reconstruction loss. For the v-th view of the i-th sample, we use a negative exponential function to convert the error into a target quality score:

$$\mathbf{r}_{i,s}^{(v)} = \exp(-||\mathbf{x}_i^{(v)} - \hat{\mathbf{x}}_i^{(v)}||_2^2). \tag{11}$$

Then we normalize the target quality scores such that their sum is equal to 1:

$$\hat{\mathbf{r}}_{i,s}^{(v)} = \frac{\mathbf{r}_{i,s}^{(v)}}{\sum_{v' \in \mathcal{V}_i} \mathbf{r}_{i,s}^{(v')} + \epsilon}. \tag{12}$$

Finally, the ultimate quality discrimination loss is obtained by calculating the Mean Squared Error between the quality scores given by the quality assessment network and the target quality scores:

$$\mathcal{L}_{qa} = \frac{1}{\sum_{i=1}^{\mathbf{n}} |\mathcal{V}_i|} \sum_{i=1}^{n} \sum_{v \in \mathcal{V}_i} ||\hat{\mathbf{q}}_{i,s}^{(v)} - \hat{\mathbf{r}}_{i,s}^{(v)}||_2^2. \tag{13}$$

Multi-label Classification and Objective Function

After obtaining the fused multi-view distribution $\mathcal{N}(\hat{\mu}_{i,s}, (\hat{\sigma}_{i,s})^2)$, we utilize the reparameterization trick to sample from it, thus acquiring the multi-view fused representation $\bar{\mathbf{z}}$, which can be used for downstream tasks. Then, we input the fused representation into a multi-label classifier to produce the final prediction vector $\mathbf{p}$. The cross-entropy loss for multi-label classification is as follows:

$$\mathcal{L}_{ce} = -\frac{1}{N} \sum_{i=1}^{\mathbf{n}} \sum_{j \in \mathcal{G}_i} [(1 - \mathbf{y}_{ij}) \log(1 - \mathbf{p}_{ij}) + \mathbf{y}_{ij} \log \mathbf{p}_{ij}]. \tag{14}$$

where $N = \sum_{i=1}^{\mathbf{n}} |\mathcal{G}_i|$ and $\mathbf{p}_{i,j}$ and $\mathbf{y}_{i,j}$ denote the prediction and the ground-truth label of the j-th class for the i-th sample, respectively. Finally, the overall optimization objective of our model is composed of the loss terms from the aforementioned modules. Its specific form is as follows:

$$\mathcal{L} = \mathcal{L}_{dis} + \mathcal{L}_{sc} + \beta \mathcal{L}_{qa} + \mathcal{L}_{ce}. \tag{15}$$

where β is a balancing factor. The Algorithm 1 illustrates the training process for our proposed method.

Algorithm 1: Training process of our model

Input: Incomplete multi-view data $\{\{\mathbf{x}_i^{(v)}\}_{v=1}^{\mathbf{m}}\}_{i=1}^{\mathbf{n}}$, known view set $\{\mathcal{V}_i\}_{i=1}^{\mathbf{n}}$, incomplete multi-label $\mathbf{Y}$, known label set $\{\mathcal{G}_i\}_{i=1}^{\mathbf{n}}$.
Initialization: Initialize the parameters of the model and set hyper-parameters (α, β, γ and training epochs $\mathbf{E}$).
Output: Trained parameters of the model.

for $e = 1, ..., \mathbf{E}$ **do**
  Disentangle the available views using the view encoder set $\{E_s^{(v)}, E_p^{(v)}\}_{v=1}^{\mathbf{m}}$ to obtain the shared and private representation distributions.
  Compute the view reconstruction loss using Eq. (1).
  Compute the KL loss of each distribution by Eq. (2).
  For the shared and private representations of each view, calculate the orthogonality loss using Eq. (3).
  Fuse the distribution parameters based on the weights output by the quality assessment network.
  Sample the representation $\bar{\mathbf{z}}$ from the fused distribution, then feed $\bar{\mathbf{z}}$ into the multi-label classifier to obtain the prediction $\mathbf{p}$.
  Compute the consistency constraint loss by Eq. (7).
  Compute the quality discrimination loss by Eq. (13).
  Compute the classification loss using Eq. (14).
  Calculate the total loss $\mathcal{L}$ using Eq. (15).
  Update the parameters of model batch to batch.
end for

Experiments

Experimental Settings

Datasets: To evaluate the performance of our model and ensure a fair comparison with existing state-of-the-art methods, we conducted experiments on five benchmark datasets that are commonly used in this field: i.e., Corel5k (Duygulu et al. 2002), Pascal07 (Everingham et al. 2010), ESPGame (Von Ahn and Dabbish 2004), IAPRTC12 (Grubinger et al. 2006) and Mirflickr (Huiskes and Lew 2008). Each sample in the datasets includes six features: GIST, HSV, DenseHue, DenseSift, RGB, and LAB. Detailed information for each dataset are provided in the Appendix A.

Incomplete multi-view missing multi-label data preprocessing: To construct the incomplete multi-view and missing multi-label datasets required for our experiments, we started with the six aforementioned standard, complete datasets. Following a strategy consistent with existing state-of-the-art methods (Tan et al. 2018; Li and Chen 2021), we artificially introduced missingness for both views and labels to simulate real-world imperfect data scenarios. Specifically, for each view, we randomly mask its data according to a predefined masking probability. However, we ensure that each sample retains at least one available view. Similarly, we randomly mask 50% of the positive and negative labels of each sample. Each dataset was randomly divided into a training set (70%), a validation set (15%), and a test set (15%).

Comparison methods: In the experimental comparison section, we selected a total of nine state-of-the-art meth-

Data	Metric	CDMM	DM2L	LVSL	iMVWL	NAIM3L	DICNet	DIMC	MSLPP	SIP	**QUALITY**
Corel5K	AP	$0.354_{0.004}$	$0.262_{0.005}$	$0.342_{0.004}$	$0.283_{0.008}$	$0.309_{0.004}$	$0.381_{0.004}$	$0.353_{0.006}$	$0.413_{0.008}$	$0.418_{0.009}$	$0.434_{0.010}$
	1-HL	$0.987_{0.000}$	$0.987_{0.000}$	$0.987_{0.000}$	$0.978_{0.000}$	$0.987_{0.000}$	$0.988_{0.000}$	$0.987_{0.000}$	$0.988_{0.000}$	$0.988_{0.000}$	$0.988_{0.000}$
	1-RL	$0.884_{0.003}$	$0.843_{0.002}$	$0.881_{0.003}$	$0.865_{0.005}$	$0.878_{0.002}$	$0.882_{0.004}$	$0.867_{0.001}$	$0.901_{0.003}$	$0.911_{0.003}$	$0.913_{0.003}$
	AUC	$0.888_{0.003}$	$0.845_{0.002}$	$0.884_{0.003}$	$0.868_{0.005}$	$0.881_{0.002}$	$0.884_{0.004}$	$0.870_{0.001}$	$0.903_{0.004}$	$0.913_{0.003}$	$0.915_{0.003}$
	1-OE	$0.410_{0.007}$	$0.295_{0.014}$	$0.391_{0.009}$	$0.311_{0.015}$	$0.350_{0.009}$	$0.468_{0.007}$	$0.422_{0.015}$	$0.485_{0.001}$	$0.489_{0.016}$	$0.511_{0.018}$
	1-Cov	$0.723_{0.007}$	$0.647_{0.005}$	$0.718_{0.006}$	$0.702_{0.008}$	$0.725_{0.005}$	$0.727_{0.011}$	$0.684_{0.011}$	$0.766_{0.009}$	$0.787_{0.009}$	$0.795_{0.008}$
Pascal07	AP	$0.508_{0.005}$	$0.471_{0.008}$	$0.504_{0.005}$	$0.437_{0.018}$	$0.488_{0.003}$	$0.505_{0.012}$	$0.532_{0.002}$	$0.544_{0.010}$	$0.555_{0.010}$	$0.590_{0.009}$
	1-HL	$0.931_{0.001}$	$0.928_{0.001}$	$0.930_{0.000}$	$0.882_{0.004}$	$0.928_{0.001}$	$0.929_{0.001}$	$0.931_{0.001}$	$0.932_{0.001}$	$0.931_{0.001}$	$0.934_{0.001}$
	1-RL	$0.812_{0.004}$	$0.761_{0.005}$	$0.806_{0.003}$	$0.736_{0.015}$	$0.783_{0.001}$	$0.783_{0.008}$	$0.813_{0.000}$	$0.819_{0.006}$	$0.830_{0.004}$	$0.852_{0.005}$
	AUC	$0.838_{0.003}$	$0.779_{0.004}$	$0.832_{0.002}$	$0.767_{0.015}$	$0.811_{0.001}$	$0.809_{0.006}$	$0.833_{0.002}$	$0.841_{0.004}$	$0.850_{0.005}$	$0.870_{0.004}$
	1-OE	$0.419_{0.008}$	$0.420_{0.011}$	$0.419_{0.008}$	$0.362_{0.023}$	$0.421_{0.006}$	$0.427_{0.015}$	$0.456_{0.011}$	$0.466_{0.014}$	$0.464_{0.018}$	$0.499_{0.013}$
	1-Cov	$0.759_{0.003}$	$0.692_{0.004}$	$0.751_{0.003}$	$0.677_{0.015}$	$0.727_{0.002}$	$0.731_{0.006}$	$0.769_{0.007}$	$0.771_{0.003}$	$0.783_{0.006}$	$0.808_{0.006}$
ESPGame	AP	$0.289_{0.003}$	$0.212_{0.002}$	$0.285_{0.003}$	$0.244_{0.005}$	$0.246_{0.002}$	$0.297_{0.002}$	$0.287_{0.002}$	$0.310_{0.004}$	$0.311_{0.004}$	$0.318_{0.004}$
	1-HL	$0.983_{0.000}$	$0.982_{0.000}$	$0.983_{0.000}$	$0.972_{0.000}$	$0.983_{0.000}$	$0.983_{0.000}$	$0.983_{0.000}$	$0.983_{0.000}$	$0.983_{0.000}$	$0.983_{0.000}$
	1-RL	$0.832_{0.001}$	$0.781_{0.001}$	$0.829_{0.001}$	$0.808_{0.002}$	$0.818_{0.002}$	$0.832_{0.001}$	$0.821_{0.000}$	$0.843_{0.002}$	$0.849_{0.002}$	$0.854_{0.002}$
	AUC	$0.836_{0.001}$	$0.785_{0.001}$	$0.833_{0.002}$	$0.813_{0.002}$	$0.824_{0.002}$	$0.836_{0.001}$	$0.826_{0.000}$	$0.847_{0.002}$	$0.853_{0.002}$	$0.858_{0.002}$
	1-OE	$0.396_{0.005}$	$0.294_{0.006}$	$0.389_{0.004}$	$0.343_{0.013}$	$0.339_{0.003}$	$0.439_{0.007}$	$0.435_{0.009}$	$0.457_{0.012}$	$0.455_{0.007}$	$0.461_{0.008}$
	1-Cov	$0.574_{0.004}$	$0.488_{0.003}$	$0.567_{0.005}$	$0.548_{0.004}$	$0.571_{0.003}$	$0.593_{0.003}$	$0.562_{0.004}$	$0.622_{0.005}$	$0.628_{0.005}$	$0.641_{0.004}$
IAPRTC12	AP	$0.305_{0.004}$	$0.234_{0.003}$	$0.304_{0.004}$	$0.237_{0.003}$	$0.261_{0.001}$	$0.323_{0.001}$	$0.308_{0.001}$	$0.340_{0.005}$	$0.331_{0.006}$	$0.347_{0.004}$
	1-HL	$0.981_{0.000}$	$0.980_{0.000}$	$0.981_{0.000}$	$0.969_{0.000}$	$0.980_{0.000}$	$0.981_{0.000}$	$0.980_{0.000}$	$0.981_{0.000}$	$0.980_{0.000}$	$0.981_{0.000}$
	1-RL	$0.862_{0.002}$	$0.823_{0.002}$	$0.861_{0.002}$	$0.833_{0.002}$	$0.848_{0.001}$	$0.873_{0.001}$	$0.864_{0.000}$	$0.882_{0.002}$	$0.885_{0.003}$	$0.893_{0.002}$
	AUC	$0.864_{0.002}$	$0.825_{0.001}$	$0.863_{0.001}$	$0.835_{0.001}$	$0.850_{0.001}$	$0.874_{0.000}$	$0.864_{0.000}$	$0.883_{0.002}$	$0.886_{0.002}$	$0.894_{0.002}$
	1-OE	$0.432_{0.008}$	$0.340_{0.006}$	$0.429_{0.009}$	$0.352_{0.008}$	$0.390_{0.005}$	$0.468_{0.002}$	$0.431_{0.006}$	$0.474_{0.008}$	$0.463_{0.009}$	$0.481_{0.007}$
	1-Cov	$0.597_{0.004}$	$0.529_{0.004}$	$0.597_{0.004}$	$0.564_{0.005}$	$0.592_{0.004}$	$0.649_{0.001}$	$0.597_{0.004}$	$0.672_{0.006}$	$0.675_{0.007}$	$0.693_{0.006}$
Mirflickr	AP	$0.570_{0.002}$	$0.514_{0.006}$	$0.553_{0.002}$	$0.490_{0.012}$	$0.551_{0.002}$	$0.589_{0.005}$	$0.602_{0.002}$	$0.615_{0.004}$	$0.614_{0.004}$	$0.633_{0.004}$
	1-HL	$0.886_{0.001}$	$0.878_{0.001}$	$0.885_{0.001}$	$0.839_{0.002}$	$0.882_{0.001}$	$0.888_{0.002}$	$0.888_{0.000}$	$0.892_{0.001}$	$0.891_{0.001}$	$0.895_{0.001}$
	1-RL	$0.856_{0.001}$	$0.831_{0.003}$	$0.856_{0.001}$	$0.803_{0.008}$	$0.844_{0.001}$	$0.863_{0.004}$	$0.865_{0.001}$	$0.879_{0.002}$	$0.877_{0.002}$	$0.886_{0.002}$
	AUC	$0.846_{0.001}$	$0.828_{0.003}$	$0.844_{0.001}$	$0.787_{0.012}$	$0.837_{0.001}$	$0.849_{0.004}$	$0.852_{0.001}$	$0.865_{0.002}$	$0.860_{0.003}$	$0.872_{0.002}$
	1-OE	$0.631_{0.004}$	$0.510_{0.008}$	$0.607_{0.004}$	$0.511_{0.022}$	$0.585_{0.003}$	$0.637_{0.007}$	$0.647_{0.007}$	$0.667_{0.007}$	$0.662_{0.008}$	$0.682_{0.006}$
	1-Cov	$0.640_{0.001}$	$0.604_{0.005}$	$0.636_{0.001}$	$0.572_{0.013}$	$0.631_{0.002}$	$0.652_{0.007}$	$0.661_{0.003}$	$0.679_{0.003}$	$0.678_{0.003}$	$0.695_{0.003}$

Table 1: Experimental results of ten methods on five datasets with 50% missing-view rate and 50% missing-label rate (the bottom right digit is the standard deviation).

ods as baselines. These include the six methods detailed in the related work section: CDMM, LVSL, DM2L, iMVWL, NAIM3L and DICNet. In addition, we include three other methods: DIMC (Wen et al. 2023), MSLPP (Long et al. 2024), and SIP (Liu et al. 2024c). Detailed of these methods are provided in the Appendix B. Baselines not directly applicable to the iM3C problem were adapted accordingly, with implementation details available in the Appendix B.

Evaluation metrics: Following established practices in the literature (Tan et al. 2018; Li and Chen 2021), we evaluate the performance of our model by six metrics: Ranking Loss (RL), Average Precision (AP), Hamming Loss (HL), Area Under the ROC Curve (AUC), One-Error (OE), and Coverage (Cov). For RL, HL, OE, and Cov, a lower value indicates better performance. To unify the evaluation, we report their results as 1-metric. Consequently, for the six metrics, a higher score signifies superior performance.

Experimental Results and Analysis

We conducted a comprehensive performance comparison of our proposed method against nine other state-of-the-art methods on five public datasets. Table 1 presents the comparative results under the experimental setting where both the view and label missing rates were 50%. The table details the performance values of all models across all six evaluation metrics. Based on the experimental results in Table 1, we draw the following key observations:

- Comprehensive experimental comparisons reveal that our proposed method significantly surpasses nine other approaches on all performance metrics across five datasets. This shows the advanced standing of the method and exceptional the ability to address the iM3C problem.
- Models tailored for the iM3C problem outperform simple adaptations, showing it is a distinct challenge that requires purpose-built algorithms.

Furthermore, to test the generality of our proposed method, we further evaluated its performance under the standard setting where the data is fully complete. We present the multi-metric comparison results between our model and the baseline methods in the form of a radar chart in Fig. 3. It is worth noting that even in the complete data scenario, our method's performance remains comprehensively superior.

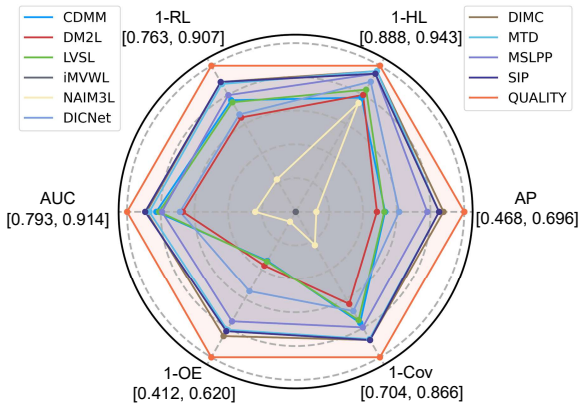


Figure 3: The experimental results of nine methods on the complete Pascal07 dataset. In this visualization, the worst results are positioned at the center.

Investigation of the Quality Assessment Network

This section provides a focused analysis of a key component, the Quality Assessment Network. A simpler alternative would be to skip the learning process of this network and directly use our supervisory proxy signal, that is, the negative exponential of the reconstruction loss as the fusion weights. In fact, we experimentally validated this simplified scheme and the results shown in Table 2. The results confirm the effectiveness of the quality assessment network. Learning the mapping with a dedicated network, rather than using a fixed strategy of converting reconstruction loss to weights, yields a smoother, more reliable quality assessment and ultimately, superior performance.

Strategies	Dataset	AP	1-HL	1-RL	AUC
Pred Fusion	Corel5K	0.434	0.988	0.913	0.915
Direct Fusion	Corel5K	0.426	0.987	0.910	0.912
Pred Fusion	Pascal07	0.590	0.934	0.852	0.870
Direct Fusion	Pascal07	0.581	0.933	0.848	0.866

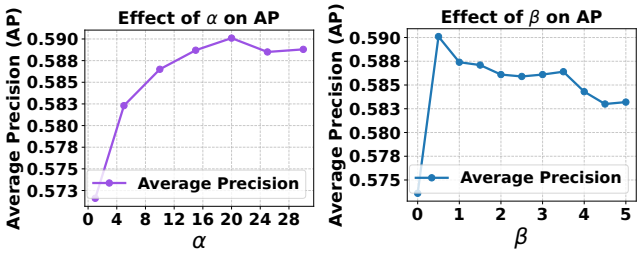
Table 2: Performance of our method under different fusion strategies, where 'Pred Fusion' uses the weights output by the quality assessment network for fusion, while 'Direct Fusion' uses the normalized negative exp-loss for fusion.

Ablation Study

To validate the effectiveness of each component of our model, we conducted ablation studies on the Corel5k and Pascal07 datasets. Besides the conventional reconstruction and classification losses, the innovation of our model is embodied in two new key constraints: the shared soft consistency constraint loss $\mathcal{L}_{sc}$ and the quality assessment loss $\mathcal{L}_{qa}$. Therefore, our ablation study will primarily focus on the effectiveness of these. As presented in Table 3, the ablation study results provide strong evidence that both constraints are integral and effective parts of our model. Furthermore, it is their synergy that yields the optimal results.

Backbone	$\mathcal{L}_{sc}$	$\mathcal{L}_{qa}$	Corel5k		Pascal07	
			AP	AUC	AP	AUC
✓			0.3406	0.8801	0.5544	0.8522
✓	✓		0.3691	0.8906	0.5736	0.8621
✓		✓	0.4057	0.9081	0.5708	0.8605
✓	✓	✓	0.4335	0.9153	0.5901	0.8696

Table 3: The ablation experimental results on the Corel5K and Pascal07 datasets, and all datasets are with 50% missing views, 50% missing labels and 70% training samples.



(a) AP values of test set (b) AP values of test set

Figure 4: Hyper-parameters analysis regarding α and β on the Pascal07 dataset (with a setting of 50% view and 50% label double missingness).

Hyper-parameter Sensitivity Study

In this section, we performed a sensitivity analysis of two key hyperparameters on the Pascal07 dataset. Specifically, Fig. 4(a) depicts the performance fluctuation of the model as α varies within the range of [1,30], while Fig. 4(b) correspondingly shows the results as β varies within the range of [0,5]. The experimental results show that our model has good parameter robustness: the model's performance remains stable when the values of hyperparameters α and β are within the relatively wide ranges of [15,30] and [0.5,3.5], respectively. For γ, we recommend setting it to $1e-3$.

Conclusion

In this paper, we first analyze the problem of information loss caused by strong consistency constraints on shared representations in the context of multi-view shared and private representation disentanglement. Based on this, we propose a framework named **QUALITY** for the iM3C task. On the one hand, through the soft consistency constraint brought about by weighted contrastive learning, we not only align the shared representations but also enhance the aggregation effect of high-quality representations on other representations. On the other hand, we designed a quality assessment network to evaluate the representations, thus achieving instance-level dynamic fusion. Finally, extensive experimental results on multiple datasets have verified the effectiveness and robustness of our method.

Acknowledgments

This study was supported in part by the Shenzhen Science and Technology Program (Grant No. JCYJ20240813105135047), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515030213), and in part by the Laboratory for Artificial Intelligence in Design, the InnoHK Initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government.

References

Chen, H.; Qu, C.; Zhang, Y.; Chen, C.; and Jiao, J. 2023. Multi-view self-supervised disentanglement for general image denoising. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12281–12291.

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 *Computer vision—ECCV 2002: 7th European conference on computer vision copenhagen, Denmark, May 28–31, 2002 proceedings, part IV 7*, 97–112. Springer.

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 ontoImage*, volume 2.

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.

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

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.

Li, X.; Pan, Y. P.; Sun, Y.; Sun, Q.; Sun, Y.; W. Tsang, I.; and Ren, Z. 2025. Incomplete Multi-view Clustering with Paired and Balanced Dynamic Anchor Learning. *IEEE Transactions on Multimedia*, 7087–7098.

Li, X.; Pan, Y. P.; Sun, Y.; Sun, Q. S.; Tsang, I. W.; and Ren, Z. 2024. Fast Unpaired Multi-view Clustering.

Lin, Y.; Dou, X.; Luo, X.; Wu, Z.; Liu, C.; Luo, T.; Wen, J.; Ling, B. W.-k.; Xu, Y.; and Wang, W. 2025. Multi-view diabetic retinopathy grading via cross-view spatial alignment and adaptive vessel reinforcing. *Pattern Recognition*, 164: 111487.

Liu, C.; Jia, J.; Wen, J.; Liu, Y.; Luo, X.; Huang, C.; and Xu, Y. 2024a. Attention-induced embedding imputation for incomplete multi-view partial multi-label classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 13864–13872.

Liu, C.; Li, R.; Che, H.; Leung, M.-F.; Wu, S.; Yu, Z.; and Wong, H.-S. 2024b. Latent structure-aware view recovery for incomplete multi-view clustering. *IEEE transactions on knowledge and data engineering*.

Liu, C.; Wen, J.; Liu, Y.; Huang, C.; Wu, Z.; Luo, X.; and Xu, Y. 2023a. Masked two-channel decoupling framework for incomplete multi-view weak multi-label learning. *Advances in neural information processing systems*, 36: 32387–32400.

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. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 8807–8815.

Liu, C.; Wen, J.; Xu, Y.; Zhang, B.; Nie, L.; and Zhang, M. 2025. Reliable representation learning for incomplete multi-view missing multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, C.; Xu, G.; Wen, J.; Liu, Y.; Huang, C.; and Xu, Y. 2024c. Partial multi-view multi-label classification via semantic invariance learning and prototype modeling. In *Forty-first international conference on machine learning*.

Long, J.; Zhang, Q.; Lu, X.; Wen, J.; Zhao, L.; and Xie, W. 2024. Multi-scale locality preserving projection for partial multi-view incomplete multi-label learning. *Neural Networks*, 180: 106748.

Lu, Y.; Lin, Y.; Yang, M.; Peng, D.; Hu, P.; and Peng, X. 2024. Decoupled contrastive multi-view clustering with high-order random walks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 14193–14201.

Luo, X.; Xu, Q.; Wu, H.; Liu, C.; Lai, Z.; and Shen, L. 2025. Like an Ophthalmologist: Dynamic Selection Driven Multi-View Learning for Diabetic Retinopathy Grading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19224–19232.

Lv, Z.; Gao, Q.; Zhang, X.; Li, Q.; and Yang, M. 2022. View-consistency learning for incomplete multiview clustering. *IEEE Transactions on Image Processing*, 31: 4790–4802.

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

Su, C.; Zheng, H.; Peng, D.; and Wang, X. 2025. DiCA: Disambiguated Contrastive Alignment for Cross-Modal Retrieval with Partial Labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 20610–20618.

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

Tsai, Y.-H. H.; Wu, Y.; Salakhutdinov, R.; and Morency, L.-P. 2020. Self-supervised learning from a multi-view perspective. *arXiv preprint arXiv:2006.05576*.

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

Wang, X.; Hu, P.; Liu, P.; and Peng, D. 2020. Deep semisupervised class-and correlation-collapsed cross-view learning. *IEEE transactions on cybernetics*, 52(3): 1588–1601.

Wang, X.; Peng, D.; Hu, P.; Gong, Y.; and Chen, Y. 2023a. Cross-domain alignment for zero-shot sketch-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(11): 7024–7035.

Wang, X.; Peng, D.; Yan, M.; and Hu, P. 2023b. Correspondence-free domain alignment for unsupervised cross-domain image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10200–10208.

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(8): 11396–11408.

Wu, S.; Zheng, Y.; Ren, Y.; He, J.; Pu, X.; Huang, S.; Hao, Z.; and He, L. 2024. Self-weighted contrastive fusion for deep multi-view clustering. *IEEE Transactions on Multimedia*, 26: 9150–9162.

Xie, M.-K.; Xiao, J.; Liu, H.-Z.; Niu, G.; Sugiyama, M.; and Huang, S.-J. 2023. Class-distribution-aware pseudo-labeling for semi-supervised multi-label learning. *Advances in Neural Information Processing Systems*, 36: 25731–25747.

Xie, W.; Lu, X.; Liu, Y.; Long, J.; Zhang, B.; Zhao, S.; and Wen, J. 2024. Uncertainty-aware pseudo-labeling and dual graph driven network for incomplete multi-view multi-label classification. In *Proceedings of the 32nd ACM international conference on multimedia*, 6656–6665.

Zeng, P.; Yang, M.; Lu, Y.; Zhang, C.; Hu, P.; and Peng, X. 2023. Semantic invariant multi-view clustering with fully incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4): 2139–2150.

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, Q.; Wei, Y.; Han, Z.; Fu, H.; Peng, X.; Deng, C.; Hu, Q.; Xu, C.; Wen, J.; Hu, D.; et al. 2024. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*.

Zhao, D.; Gao, Q.; Lu, Y.; and Sun, D. 2022. Non-aligned multi-view multi-label classification via learning view-specific labels. *IEEE Transactions on Multimedia*, 25: 7235–7247.

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