EXOTIC: External Vision Guided Incomplete Multi-view Classification

Shilin Xu[1], Dezhong Peng[1, 3], Zhenwen Ren[2], Yuan Sun[1*]
[1]Sichuan University, Chengdu, China 610044
[2]Southwest University of Science and Technology, Mianyang, China 621010
[3]Tianfu Jincheng Laboratory, Chengdu, China 610093
xushilin990@gmail.com, pengdz@scu.edu.cn, rzw@njust.edu.cn, sunyuan_work@163.com

Abstract

Due to sensor failures and occlusions during data acquisition, multi-view data often suffer from partial missing samples, thereby producing incomplete multi-view data. Recently, Incomplete Multi-View Classification (IMVC) has become one of the research hot topics, where numerous IMVC methods have been proposed. Although these methods have achieved promising performance by exploiting internal semantic information from partially observed data, they primarily rely on limited internal supervision for view completion. Clearly, this largely constrains their performance ceiling. To overcome this limitation, we propose an EXternal visiOn-driven incomplete mulTi-vIew Classification (EXOTIC) paradigm that incorporates external vision knowledge as semantic guidance, thereby assisting in imputing incomplete views. To the best of our knowledge, it is the first work that leverages external vision knowledge as supervision signals, thereby guiding missing-view completion. Specifically, we first introduce an external vision knowledge library based on a pre-trained vision–language model. Then, we design a Knowledge Filtering module to adaptively select task-relevant knowledge. Afterwards, we present a Knowledge Purification module to align external knowledge with internal representations. Finally, we propose External Completion that leverages the refined knowledge to impute missing views, thereby enhancing the classification decision ability. Extensive experiments on multiple incomplete multi-view datasets demonstrate that the proposed EXOTIC consistently outperforms existing methods, especially under high missing rates. The code is available at: https://github.com/sstaree/EXOTIC.

1. Introduction

The proliferation of multimedia technologies has fundamentally transformed data acquisition paradigms, enabling comprehensive characterization of real-world objects through multi-view data. These data originate from diverse sources (e.g., sensors and modalities) or feature extraction methodologies, inherently exhibiting heterogeneous properties that reflect different perspectives of the same semantic content. For instance, a news story may simultaneously exist as textual articles, photographic evidence, video footage, and social media commentary across multiple languages. Each view offers unique discriminative information while collectively forming a cohesive narrative. As a result, effective integration and analysis of multi-view data have become crucial for improving performance and robustness in many machine learning tasks [7, 18, 30, 31, 33]. Benefiting from the inherent diversity and complementarity of multi-view data, multi-view learning is extensively applied in multiple fields, including clus-

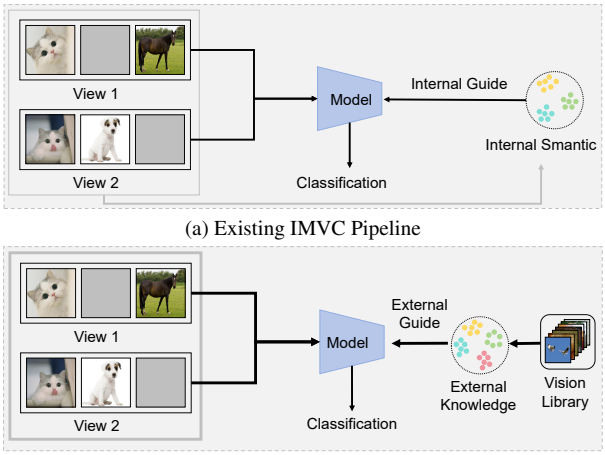


Figure 1. Our key idea. (a) Existing IMVC methods mainly focus on mining internal supervision implicit in incomplete multi-view data, which may struggle to obtain sufficient semantic information, causing the inherent performance ceiling. (b) Our intuitive idea is to incorporate external vision knowledge to enrich the semantic information, thereby enhancing the classification performance.

*Corresponding authors.

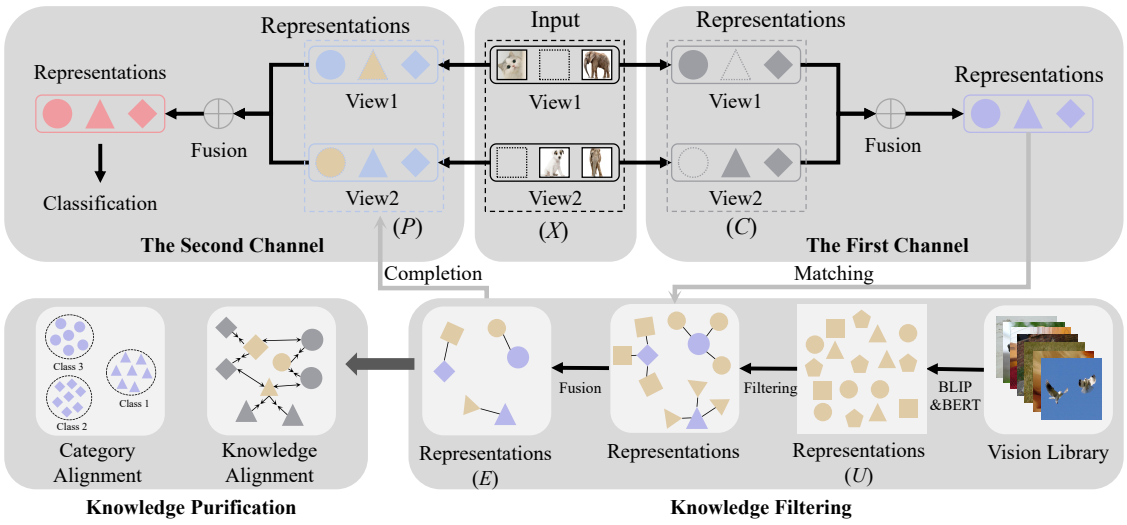


Figure 2. The framework of the proposed EXOTIC. EXOTIC is mainly composed of four key components, i.e., External Knowledge Construction, Knowledge Filtering, Knowledge Purification, and External Completion. Knowledge Filtering and Knowledge Purification are performed in the first channel, corresponding to the right channel in the diagram, while External Completion and the classification task take place in the second channel, represented by the left channel in the diagram.

tering [3–5, 10, 19, 20, 44, 47], classification [8, 9, 11, 28], and retrieval [2, 6, 15, 32, 36, 49].

Over the past years, to mitigate the information sparsity or bias limitations of single-view learning [17], a large number of Multi-View Classification (MVC) methods [9, 21, 38] have been proposed. MVC aims to effectively integrate complementary information from multiple views, thereby improving classification decision performance and generalization ability. For instance, SMDC [27] builds a safe multi-view model based on evidence theory to address performance degradation when additional views are incorporated. To ensure reliable decisions for low-quality conflictive instances, RCML [39] presents a conflictive opinion aggregation scheme that explicitly models the relationship between shared and view-specific reliabilities across multiple views. However, in the real world, multi-view data is often incomplete due to some limitations in data acquisition, sensor failures, or occlusions. This incompleteness poses a new significant challenge to multi-view learning. In the case of missing views, the key core is that how to effectively exploit the available partial information to accurately explore the complex inter-view dependencies, which directly affects the robustness and generalization ability of downstream tasks. Thus, numerous Incomplete Multi-View Clustering (IMVC) methods have been proposed, which could be roughly divided into two categories, i.e., missing imputation [43] and missing imputation-free [24]. The former leverages latent graphs, prototypes, or autoencoders to infer missing data, thereby improving classification performance. The latter primarily focuses on learning from available views to enhance consistency and complementarity among existing views.

Although existing IMVC methods [12, 41] achieve impressive performance, they suffer from an inherent performance upper limitation on semantic guidance due to relying solely on mining internal semantic information in incomplete multi-view data to learn discriminative representations. Regrettably, some valuable external knowledge that could facilitate classification has been inadvertently ignored. Recently, some studies [29] integrate external textual knowledge guidance (i.e., WordNet) into the unsupervised tasks (e.g., clustering), thereby exploring the incorporation of rich external knowledge rather than extracting internal semantics solely from the data itself, such as SIC [1] and TAC [16]. Since external knowledge could provide more supervisory signals through the semantic information contained in textual descriptions, these methods could significantly improve the performance of unsupervised tasks. However, external knowledge often comes from heterogeneous sources, with inconsistent semantic bias. Thus, in supervised learning scenarios, when external knowledge deviates from the data label or the internal representation of the model, it is easy to introduce potential semantic conflict (i.e., semantic noise). This could destroy the original label distribution and weaken the discriminative ability of the model, thereby affecting the final decision performance. Therefore, **one of the key challenges lies in how to effectively reconcile internal and external knowledge to prevent semantic conflicts and ensure consistent semantic guidance**. In addition, existing external knowledge methods generally rely on large-scale textual descriptions, such as entity descriptions, attribute labels, or language model

embeddings. However, in practice, high-quality and structured textual descriptions are often difficult to obtain or have insufficient coverage. In contrast, image resources are widely distributed on the internet and in application environments, which have rich visual semantic information. However, **how to construct the vision knowledge library into a usable external knowledge to impute incomplete multi-view data still remains a significant challenge**.

To enable effective inference of missing multi-view data, we propose a novel EXternal visiOn-driven incomplete mulTi-vIew Classification (EXOTIC) paradigm to enrich the semantic guidance in incomplete multi-view learning. As shown in Fig.1, different from previous IMVC methods that focus solely on exhaustively exploring and utilizing internal supervision signals, our EXOTIC incorporates external vision knowledge to assist the completion process for incomplete multi-view data. Specifically, as shown in Fig.2, we first introduce a diverse collection of unlabeled images as an external vision knowledge library, from which deep feature representations are extracted using the pre-trained vision–language model. Then, we propose a knowledge filtering module to select the most relevant external knowledge for multi-view data, thereby mitigating the semantic ambiguity caused by noise and irrelevant distractions inherent in the unlabeled vision library. Moreover, to prevent task interference between knowledge filtering and representation learning, we design a dual-channel decoupling module to effectively decouple the two tasks. That is, one channel specializes in filtering and selecting external knowledge, while the other channel focuses on encoding the available multi-view data into high-quality embeddings and using the selected knowledge to impute missing views. Since the differences between domain-specific vision knowledge and multi-view data could cause knowledge conflicts, we propose a knowledge purification module to maximize the consistency between the selected external knowledge and the view-specific features, thereby balancing and aligning internal and external information. Finally, we use the cross-entropy loss to achieve the final prediction. Overall, the key contributions of this paper are summarized as follows

- We propose a novel external Vision-driven Incomplete Multi-view Classification (EXOTIC) method, which incorporates external vision knowledge guidance to impute missing views, thereby breaking through the inherent limitations of the internal multi-view structure. To the best of our knowledge, it is the first work to incorporate external vision knowledge into supervision signals, thereby enhancing the semantic guidance for facilitating IMVC.
- We present dual-channel decoupling to avoid task interference between knowledge filtering and representation learning. Further, to mitigate knowledge conflicts caused by external knowledge, we propose knowledge purification to enhance the consistency between knowledge and multi-view representation.
- We conduct extensive experiments on seven multi-view datasets with different missing rates. The experiments show that the proposed EXOTIC outperforms the state-of-the-art IMVC methods, especially in high missing rate scenarios.

2. Related Work

Multi-view learning primarily focuses on exploiting complementary and specific information across multiple views, thereby learning discriminative representations, which has demonstrated remarkable effectiveness in various downstream applications, particularly multi-view classification [45, 48]. For example, DCCAE [34] pioneered the use of deep autoencoders to extract compact yet informative representations while capturing complex cross-view dependencies. However, it lacks the capability to measure uncertainty in representation. To this end, ECML [39] proposes a conflict-aware opinion aggregation strategy based on uncertainty estimation to alleviate inconsistencies or conflicts across views. These methods implicitly assume that the acquired labels are completely accurate. However, in real-world scenarios, label annotations are noisy due to a variety of factors, such as human annotation errors, task ambiguity, and so on. Moreover, obtaining high-quality labels is typically expensive and time-consuming, further leading to the prevalence of noisy labels. To tackle this issue, TMNR [40] jointly refines noisy labels and estimates their associated uncertainty, which helps reduce the adverse effects of noisy labels. However, TMNR fails to correct mislabeled instances effectively. To address this limitation, NLC [42] further proposes an enhanced approach that combines neighbor-based KL divergence with a view-level mixup strategy to identify and rectify mislabeled samples robustly.

All the above methods assume that the views are complete. However, in real-world scenarios, data integrity across multiple views can hardly be guaranteed due to sensor failures or storage medium corruption. Therefore, it is necessary to investigate incomplete multi-view learning methodologies. Based on the approaches to handling missing views, incomplete multi-view learning can be divided into two main lines: i) Missing imputation methods. These methods [13, 43] aim to handle incomplete multi-view data by filling in the missing features. For instance, For instance, AIMNet [23] introduces an attention-driven mechanism for imputing missing instances. ii) Missing imputation-free methods. These methods [21, 24, 35] use available views and directly learn the common latent sub-space or representation for all views. For instance, LMVCAT [22] designs a masked view-aware self-attention module to handle incomplete views. Although these methods have achieved promising results, they heavily rely on the internal information of

the existing views, while ignore rich and available external knowledge. If we can fully exploit the useful discriminative information contained in external knowledge, then external information can significantly improve the discriminative performance of the model and improve the classification accuracy. Notably, a growing body of research has explored leveraging external knowledge to enhance the performance of original tasks. For instance, Visual Table [50] creates the visual table by LLMs to offer enriched multi-scale information, leading to enhanced performance on downstream tasks and TAC [16] introduces taking advantage of external knowledge to facilitate clustering and achieve impressive performance. These methods integrate external textual knowledge guidance (i.e., WordNet) into the unsupervised tasks (e.g., clustering). However, high-quality and structured textual descriptions are often difficult to obtain or have insufficient coverage. To this end, we introduce a vision library as the source of external knowledge, which is easy to obtain and contains rich semantic information.

3. Method

3.1. Overview

Let $\left\{\boldsymbol{X}^v \in \mathbb{R}^{N\times d_v}\right\}_{v=1}^{V}$ denote a multi-view dataset with N instances, where V is the number of views and d_v is the feature dimensionality of the v-th view. The class label of the i-th instance is represented by a one-hot vector $y_i \in \{0,1\}^N$, and C denotes the total number of classes. To model incomplete multi-view data, we introduce a missing-view indicator matrix $\mathcal{W} \in \{0,1\}^{N\times V}$, where $\mathcal{W}_{i,j} = 1$ indicates that the j-th view of the i-th instance is available, and $\mathcal{W}_{i,j} = 0$ otherwise. For notational simplicity, missing views are filled with zeros. Our objective is to learn a classification model from such incomplete multi-view instances to achieve accurate class prediction for unlabeled incomplete test data.

Generally, the central challenge of IMVC lies in constructing effective semantic guidance from incomplete multi-view data to achieve reliable inference of missing views. However, the semantic information that can be extracted directly from the data is inherently constrained by the multi-view structure. Recent approaches [1, 16] attempt to address this issue by introducing external knowledge, which provides richer semantic guidance to enhance model performance. Despite their potential, these methods encounter three key challenges: (1) constructing accessible and high-quality external knowledge while filtering out noise and irrelevant information; (2) reconciling internal and external knowledge to avoid semantic conflicts; and (3) effectively leveraging external knowledge to complete the missing views.

To this end, we propose an external vision-driven incomplete multi-view classification paradigm to enhance the semantic guidance, thereby enabling robust multi-view learning in scenarios with missing views. Our EXOTIC mainly consists of four key components, i.e., External Knowledge Construction, Knowledge Filtering, Knowledge Purification, and External Completion. Specifically, we first construct an external vision knowledge library composed of a diverse, unlabeled image collection. From this library, we extract deep feature representations using a pre-trained vision–language model, thereby capturing rich semantic priors without requiring manual annotation. Then, we present a knowledge filtering module to ensure relevance and reduce distraction from noisy or semantically unrelated samples. This module dynamically selects the most pertinent external knowledge tailored to each multi-view instance, effectively suppressing irrelevant or ambiguous signals inherent in the raw vision library. Furthermore, to reconcile potential semantic conflict between domain-specific multi-view data and general external vision knowledge, we design a Knowledge Purification module, which could maximize feature-level consistency between the selected external knowledge and view-specific representations. Finally, we utilize External Completion to fill the missing multi-view data, thereby achieving the final category predictions. In general, our total loss function can be expressed as

$$\mathcal{L}_{all} = \mathcal{L}_c + \alpha \cdot \mathcal{L}_{kc} + \beta \cdot \mathcal{L}_{vc} \tag{1}$$

where α and β are corresponding penalty parameters. $\mathcal{L}_c$ is the cross-entropy classification loss. And $\mathcal{L}_{vc}$ and $\mathcal{L}_{kc}$ are the category alignment loss and knowledge alignment loss in the knowledge purification module, respectively.

3.2. External Knowledge Construction

To obtain more semantic information, we choose a diverse and unlabeled image collection as the general external knowledge, such as a subset of ImageNet, which includes 50,000 pictures from 1,000 categories. We input the unlabeled image collection into the BLIP model [14] to generate textual descriptions for each image, and then use BERT to convert these descriptions into representations. The transformation from visual signals to language provides high-level semantic information that raw pixel data can not provide. Then, these textual descriptions are encoded into vector representations M. In this paper, we design an encoder $\{E_u : M \to U\}$ to further extract the external knowledge representations obtained from the BLIP model, facilitating the subsequent filtering and purification processes.

3.3. Knowledge Filtering

Although the constructed external vision library contains abundant semantic information, the direct use of all retrieved knowledge may introduce substantial noise and semantic bias, thereby undermining the reliability of view

completion. To ensure that only task-relevant and discriminative external knowledge contributes to the reconstruction of missing views, we design a Knowledge Filtering module that adaptively selects semantically aligned external knowledge for each instance.

To reduce task interference between knowledge filtering and representation learning, we introduce a dual-channel decoupling architecture in EXOTIC. The first channel $\{E_c^v : X^v \to C^v\}_{v=1}^V$ focuses on filtering and refining external knowledge, while the second channel $\left\{E_p^v : X^v \to P^v\right\}_{v=1}^V$ is dedicated to representation learning and classification. This design enables the first channel to concentrate on extracting task-relevant external knowledge, and the second channel to leverage the purified knowledge for accurate imputation, thus achieving learning without mutual interference. We define representation C and representation P as $\{C^v\}_{v=1}^V$ and $\{P^v\}_{v=1}^V$, respectively.

For each sample x_i, the first-channel encoder E_c^v produces view-specific representations $c_i^{(v)} \in C^v$, which are fused into a unified internal representation

$$z_i = \sum_{v=1}^{V} c_i^{(v)} \mathcal{W}_{i,v} / \sum_{v=1}^{V} \mathcal{W}_{i,v} \tag{2}$$

where $\mathcal{W}_{i,v}$ indicates whether the v-th view of the i-th sample exists. We then compute the cosine similarity between the internal representation z_j and each external knowledge embedding $u_i \in U$ to measure semantic affinity

$$S\left(u_i, z_j\right) = \frac{u_i (z_j)^\top}{\|u_i\| \, \|z_j\|} \tag{3}$$

where u_i denotes the representation of the i-th external knowledge representation from U, and z_j is the fusion representation of the first channel for j-th sample. Then, we select the top K external knowledge representations $\{u_{i,k}\}_{k=1}^K$ for sample i, where $u_{i,k}$ denotes the k-th filtered knowledge of sample i. Finally, we perform a average fusion to obtain the sample-specific external knowledge by

$$e_i = \frac{1}{K} \sum_{k=1}^{K} u_{i,k} \tag{4}$$

where e_i is sample-specific external knowledge of i-th sample, which is used to impute the missing views in the second channel. Through this filtering operation, we can preliminarily filter relevant external knowledge for each sample. However, these external knowledge representations still contain a large amount of task-irrelevant and noisy information, which requires further purification.

3.4. Knowledge Purification

Through the previous filtering operation, we generate a piece of external knowledge for each sample. However, these external knowledge cannot be directly utilized to complete the missing views due to the presence of noise and task-irrelevant information. To address this issue, we design two losses to purify the external knowledge. First, to ensure that the filtered external knowledge is highly relevant to the task labels, we adopt the category alignment loss as follows

$$\mathcal{L}_{vc} = -log \left(\frac{\sum_{i=1}^{N} \sum_{j \neq i}^{N} \cdot \mathbb{I}_{[\gamma]} \cdot \exp\left(\frac{S(z_i, z_j)}{\tau}\right)}{\sum_{i=1}^{N} \sum_{j \neq i}^{N} \exp\left(\frac{S(z_i, z_j)}{\tau}\right)} \right) \tag{5}$$

where $\mathbb{I}_{[\gamma]}$ is the indicator function, which equals 1 if the labels of z_i and z_j are identical and 0 otherwise. Guided by this loss, we can ensure that the external knowledge filtered by the first channel remains task-relevant to specific classification tasks.

To ensure consistency between external knowledge and internal representations while maintaining a balanced integration of internal and external information, we design a knowledge alignment loss as follows

$$\mathcal{L}_{kc} = \sum_{v=1}^{V} \frac{1}{\log(q+1)} \left(1 - \frac{\sum_{i=1}^{N} \mathcal{W}_{i,v} \cdot \exp\left(\frac{S(e_i, c_i^v)}{\tau}\right)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \mathcal{W}_{i,v} \cdot \exp\left(\frac{S(e_i, c_j^v)}{\tau}\right)} \right)^q \tag{6}$$

where $S(\cdot)$ denotes the cosine similarity function, while q and τ is the balance parameter. e_i is the external knowledge of the i-th sample, and $c_i^v \in C^v$ is the view-specific representation of the v-th view of sample i. Through the optimization of this loss function, we facilitate the alignment between external knowledge and view-specific representations, thereby enhancing the coherence and relevance of external knowledge.

3.5. External Completion

To achieve effective view reconstruction while avoiding task interference, we adopt the dual-channel decoupling strategy, where the two channels collaborate but serve distinct purposes. Specifically, the first channel is dedicated to knowledge filtering and purification, ensuring that only task-relevant is retained and balances internal and external semantic information. The second channel focuses on representation learning and classification, leveraging the refined external knowledge from the first channel to complete missing views. This decoupled design not only reduces optimization interference between different tasks but also enhances training stability and the quality of representation.

For each sample x_i, the second-channel encoder E_p^v produces view-specific representations $p_i^{(v)} \in P^v$. The purified knowledge embedding e_i obtained from the first channel is incorporated into the second channel to recover the missing views. The process of completion and fusion representation of second channel is formulated as

$$\hat{p}_i = \sum_{v=1}^{V} \left(\mathcal{W}_{i,v} \cdot p_i^v + (1 - \mathcal{W}_{i,v}) \cdot e_i \right) \tag{7}$$

Table 1. Classification accuracy (%) of our EXOTIC and ten compared methods on all datasets with different missing rates.

MR	Method	Caltech101	Scene15	LandUse21	HW	Fashion	NUSWIDE	CUB
0.1	CPM-NET	80.17±1.68	31.88±1.34	21.38±0.70	30.75±3.31	75.06±0.37	26.20±0.56	73.75±0.42
	DCP	85.19±3.78	75.94±1.17	72.14±1.54	95.90±1.74	96.15±0.64	31.76±1.91	87.00±2.61
	UIMC	90.06±0.70	66.71±1.11	52.19±0.90	74.75±2.40	95.71±0.21	30.27±0.38	45.83±4.68
	DIMC	89.06±1.96	77.28±1.59	70.52±1.69	97.60±0.51	97.29±0.39	46.80±0.23	78.75±6.25
	DICNET	89.10±1.97	76.43±1.33	70.52±1.88	97.55±0.56	97.52±0.28	46.82±0.25	87.17±4.07
	LMVCAT	91.41±0.10	76.16±1.62	65.90±3.20	98.50±0.00	87.50±0.80	42.96±0.00	88.33±0.00
	MTD	40.06±1.35	46.23±1.19	53.02±2.57	95.62±0.12	62.55±3.80	15.36±0.27	80.42±1.25
	AIMNET	91.10±0.21	68.66±1.86	52.44±1.05	97.75±0.50	92.20±0.10	43.46±0.15	**90.00±2.50**
	SIP	63.15±0.21	17.41±0.44	37.26±0.12	83.25±0.25	38.20±0.00	5.23±0.27	78.75±4.58
	RANK	54.45±0.62	41.30±1.39	54.05±0.24	89.62±1.12	53.42±1.08	13.29±0.65	84.17±0.00
	EXOTIC (Our)	**94.33±0.38**	**80.98±1.71**	**80.00±2.15**	**99.00±0.42**	**97.61±0.21**	**53.26±0.22**	89.50±1.94
0.3	CPM-NET	76.23±1.40	28.92±1.29	16.90±0.62	22.85±2.34	72.54±0.95	24.19±0.05	62.50±0.83
	DCP	80.46±3.71	73.33±1.26	67.57±2.25	96.00±0.65	94.77±0.37	30.58±2.13	83.17±1.33
	UIMC	81.13±1.99	64.68±0.51	54.52±2.06	78.06±3.11	92.47±0.68	29.71±0.17	57.71±4.73
	DIMC	83.31±2.36	72.29±0.58	69.76±2.66	95.80±0.62	95.52±0.38	48.37±0.38	78.75±5.42
	DICNET	83.31±2.23	71.84±1.12	69.05±2.87	96.00±0.50	95.56±0.54	48.37±0.47	86.83±3.89
	LMVCAT	85.92±0.21	67.95±2.13	56.25±1.32	96.62±1.12	88.65±0.60	44.69±0.12	87.92±1.25
	MTD	30.95±4.45	46.23±1.19	36.63±1.76	89.62±1.38	58.08±2.22	13.78±0.02	73.33±2.50
	AIMNET	83.64±0.41	60.09±2.31	46.73±2.00	95.38±0.12	91.47±0.17	49.15±0.59	85.83±0.00
	SIP	57.04±3.21	19.29±2.44	36.31±0.60	77.38±1.62	38.08±4.57	7.30±0.35	65.00±0.83
	RANK	45.76±0.83	40.85±1.39	34.17±0.83	87.50±0.00	49.53±1.68	11.65±0.11	82.08±2.92
	EXOTIC (Our)	**90.97±0.83**	**76.23±1.31**	**77.57±2.25**	**98.30±0.75**	**96.99±0.24**	**57.30±0.51**	**90.50±2.51**
0.5	CPM-NET	77.32±1.28	31.91±1.36	17.00±1.42	23.40±3.27	74.46±1.06	26.32±0.35	64.58±2.08
	DCP	74.52±4.61	65.66±0.63	57.57±1.73	94.30±0.48	92.01±1.29	28.23±1.43	78.33±6.12
	UIMC	71.45±1.79	59.42±1.24	45.24±2.73	77.38±2.22	89.57±1.05	28.24±0.31	46.04±5.01
	DIMC	81.47±1.78	67.76±1.29	59.81±2.31	95.35±0.75	93.67±0.53	43.29±0.38	74.17±4.17
	DICNET	81.38±1.74	66.98±1.59	60.86±2.45	95.30±1.57	93.73±0.58	43.34±0.50	77.50±4.08
	LMVCAT	82.92±0.72	62.26±0.92	47.75±3.53	92.00±0.25	90.03±1.08	40.90±0.29	86.25±0.42
	MTD	39.03±2.38	50.16±2.27	38.95±3.25	86.00±0.75	63.62±1.12	11.92±0.08	62.92±2.08
	AIMNET	83.23±0.62	57.71±1.29	38.73±2.52	92.62±0.88	94.72±0.38	43.11±0.47	78.75±4.58
	SIP	55.90±1.66	18.68±0.28	38.10±0.24	71.12±0.88	37.03±3.12	6.96±0.32	52.50±2.50
	RANK	45.76±0.83	33.26±4.77	34.88±1.70	87.25±0.00	49.22±2.02	11.28±0.14	75.83±1.67
	EXOTIC (Our)	**89.61±0.77**	**73.02±1.80**	**67.57±1.85**	**96.90±0.86**	**96.90±0.30**	**52.32±0.18**	**86.33±3.06**
0.7	CPM-NET	77.82±1.05	36.03±1.58	17.62±1.21	19.45±3.16	74.56±1.10	26.62±0.18	60.00±5.00
	DCP	70.21±2.83	52.75±0.92	40.29±1.62	92.95±1.53	86.45±0.48	27.92±0.85	69.67±5.93
	UIMC	51.53±2.42	30.35±2.30	22.24±1.98	78.75±1.52	86.12±1.41	26.44±1.43	41.25±5.22
	DIMC	79.45±0.62	64.39±1.59	54.14±3.43	93.15±0.82	91.23±0.39	40.17±0.28	70.42±5.42
	DICNET	79.66±0.58	65.15±2.27	53.71±3.59	93.75±0.63	91.50±0.49	40.08±0.45	72.33±1.22
	LMVCAT	82.19±1.45	61.12±2.14	46.03±1.70	90.00±0.75	89.80±0.10	39.73±0.05	82.08±0.42
	MTD	20.70±0.41	48.74±2.60	34.70±1.55	81.50±0.25	57.68±2.88	12.81±0.13	57.92±2.92
	AIMNET	81.16±0.62	54.55±2.55	37.65±1.52	92.00±0.50	94.17±0.28	38.62±0.03	72.08±4.58
	SIP	50.41±1.14	15.24±0.06	32.14±0.24	68.00±3.75	39.10±0.60	9.16±0.10	46.25±0.42
	RANK	38.92±0.21	35.20±0.72	32.62±0.24	82.00±1.50	50.20±1.25	11.33±0.02	76.25±1.25
	EXOTIC (Our)	**88.74±0.76**	**68.85±1.59**	**62.43±1.80**	**96.40±0.77**	**95.13±0.33**	**49.25±0.38**	**83.17±3.99**
0.9	CPM-NET	78.45±0.71	36.19±1.63	17.67±1.61	19.10±5.56	76.60±1.11	26.32±0.47	64.67±0.85
	DCP	63.14±4.81	41.74±1.70	25.48±3.35	87.60±1.40	77.38±2.67	21.00±1.45	53.00±3.60
	UIMC	66.58±2.67	29.45±2.69	20.67±2.98	84.06±1.74	84.12±0.03	26.30±0.00	36.25±1.38
	DIMC	76.94±1.94	57.01±2.43	44.29±2.53	91.80±1.30	90.00±0.98	36.12±0.35	68.33±1.67
	DICNET	76.94±2.09	57.84±0.67	44.19±2.12	92.20±0.86	90.54±0.75	36.15±0.38	67.50±3.21
	LMVCAT	81.88±1.55	58.06±1.08	40.06±0.73	86.00±0.00	82.38±0.42	37.34±0.21	77.08±0.42
	MTD	20.39±0.31	46.51±1.94	38.38±1.91	76.88±0.88	56.03±0.72	18.66±0.35	56.25±0.42
	AIMNET	78.78±0.93	51.93±0.92	33.21±1.58	89.75±1.00	89.00±0.35	43.63±0.11	80.42±1.25
	SIP	43.58±1.35	12.75±1.00	29.76±1.43	63.12±1.12	29.85±3.05	9.44±0.26	43.75±2.92
	RANK	44.10±0.83	32.82±2.00	27.38±4.29	76.12±0.62	46.90±2.50	13.04±0.32	63.75±2.92
	EXOTIC (Our)	**88.16±0.65**	**62.53±0.94**	**55.24±3.16**	**94.05±1.13**	**92.55±0.60**	**44.91±0.40**	**81.00±2.91**

where $\hat{p}_i$ is the final fusion representation of sample i, which is used to generate the classification result. Finally, we employ the cross-entropy function as follows

$$\mathcal{L}_c = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p\left(\hat{p}_i\right) \tag{8}$$

where $p(\cdot)$ represents the softmax operation.

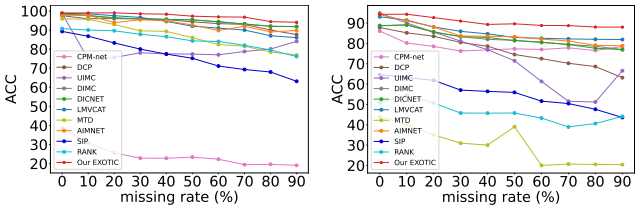
4. Experiment

4.1. Datasets and Comparison Methods

We comprehensively evaluate the classification performance of the proposed EXOTIC on seven widely used datasets. More descriptions of datasets can be found in the supplementary material[1]. To evaluate the advancement of our framework, ten excellent comparison methods are selected in our experiment, i.e., CPM-NET [46], DCP [20], UIMC [37], DIMC [35], DICNET [21], LMVCAT [22], MTD [24], AIMNET [23], SIP [25], and RANK [26]. We employ classification accuracy as the evaluation metric. We assess the robustness of EXOTIC and ten competing methods under varying Missing Rates (MR = 0.1, 0.3, 0.5, 0.7, 0.9). For fairness, we perform 5 times to record the mean accuracy and standard deviation. The best and second-best scores are highlighted by **bold** and underline, respectively.

4.2. Implementation Details

The proposed EXOTIC is implemented on an NVIDIA RTX 4080 GPU with PyTorch 2.2.2. We adopt fully connected networks with a ReLU layer to extract the view-specific representation and knowledge representation, where the dimension of the encoder is $d_{in} - 1024 - 1024 - 1500 - d_{out}$, where d_{in} is the input dimension and d_{out} is the output dimension. We employ the SGD optimizer to train our model with a batch size of 128, while the learning rates are set to 0.01 for Caltech101, HW, and Fashion datasets, 0.001 for Scene15 and NUSWIDE datasets, and 0.005 for LandUse21 dataset, respectively. We empirically set α, β, and q to 30, 1, and 0.3, respectively. The number of training epochs is set to 100.

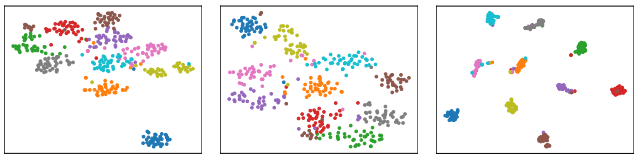


(a) HW (b) Caltech101

Figure 3. Classification performance comparisons on HW and Caltech101 datasets with different missing rates.

4.3. Results Analysis on Incomplete Data

The experimental results on all datasets with different missing rates are shown in Tab. 1. In addition, Fig. 3 presents the performance curves on the HW and Caltech101 datasets under varying missing rates. From these results, we can observe that: (1) Under low missing rate, i.e, MR = 0.1, EXOTIC already outperforms all comparison methods. (2) Under high missing rates, i.e., MR $\geq$ 0.3, EXOTIC outperforms all comparison methods on all datasets, which demonstrates the superiority of our method in handling high-missing scenarios. (3) As the missing rate increases, EXOTIC shows the smallest performance drop, which reflects the strong stability and robustness of our method in handling severe view incompleteness.



(a) Zero (b) Mean (c) EXOTIC

Figure 4. t-SNE visualization results of different completion paradigms on the HW dataset with 0.9 missing rate.

4.4. The Impact of External Knowledge

To evaluate the effectiveness of different missing view completion strategies, we conduct t-SNE visualizations to compare different completion strategies. Zero fills missing views with zeros, while Mean uses the average of existing views. As illustrated in Fig. 4, external knowledge produces clearer inter-class boundaries, indicating that it guides the model to generate more discriminative and semantically consistent representations.

Furthermore, to evaluate the impact of different types and quantities of knowledge libraries, we conduct experiments under different external knowledge libraries and quantities on seven datasets. Specifically, Leaves, Caltech101, and ImageNet serve as general natural image libraries, while iCartoonFace represents a cross-domain cartoon image library. All experiments are conducted under 0.9 missing rate. As illustrated in Tab. 2, we can observe four observations: (1) Limited external knowledge (e.g., 100 samples) leads to unstable completion performance. (2) As the quantity increases within a reasonable range, performance steadily improves and eventually saturates, demonstrating the scalability of EXOTIC. (3) Even cross-domain knowledge (iCartoonFace) achieves competitive performance, demonstrating EXOTIC's ability to leverage diverse knowledge sources.

4.5. Parameter Analysis

To evaluate the impact of each hyperparameter α, β, and q, we conduct a sensitivity analysis on the HW dataset under different missing rates. Parameter analysis of more datasets will be presented in the supplementary materials. As shown in Fig. 5, we can observe that: When α and β increase, the quality of external knowledge improves, which

[1] https://github.com/sstaree/EXOTIC

Table 2. Comparison of different external vision sources on different datasets under 0.9 missing rate.

Source	Size	Caltech101	Scene15	LandUse21	HW	Fashion	CUB	NUSWIDE
Leaves	100	75.03±11.90	63.81±0.97	44.57±6.69	65.25±32.75	91.24±0.46	80.67±2.60	44.73±0.35
Leaves	500	86.58±0.76	63.02±1.31	54.10±3.28	94.50±1.38	92.28±0.69	81.83±3.09	43.36±0.40
Caltech101	100	81.16±10.90	64.37±1.09	45.14±8.86	88.80±2.36	91.37±0.48	82.00±2.72	44.90±0.51
Caltech101	500	87.45±0.21	63.35±1.56	55.10±4.51	94.25±1.33	92.40±0.74	81.50±3.47	43.43±0.57
ImageNet	100	75.98±15.77	63.99±1.24	36.00±4.15	59.25±37.99	91.20±0.66	80.17±2.38	44.80±0.32
ImageNet	500	87.41±0.97	63.28±1.01	55.29±2.77	94.65±0.98	92.08±0.64	80.00±3.61	44.53±0.32
ImageNet	50,000	88.16±0.65	62.53±0.94	55.24±3.16	94.20±1.04	92.55±0.60	81.00±2.91	44.91±0.40
Icartoonface	2,000	87.54±0.51	63.37±0.85	55.52±3.72	94.90±1.07	92.58±0.39	80.67±2.71	43.56±0.58

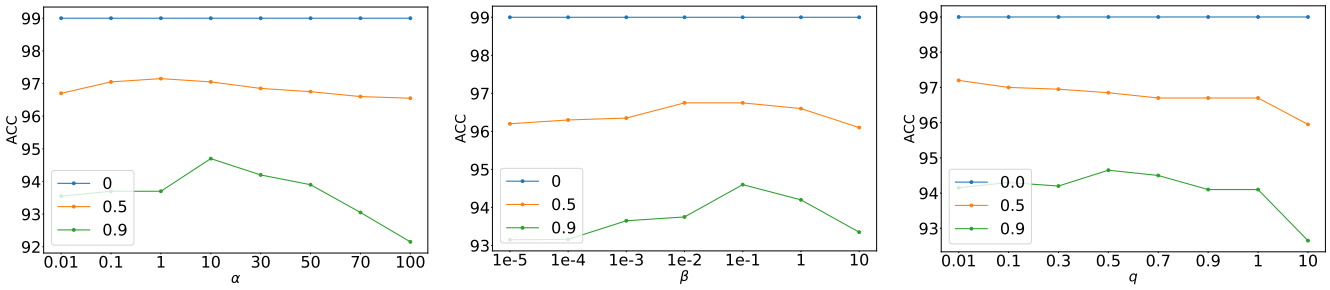


Figure 5. Parameter analysis for α, β, and q on the HW dataset with different missing rates.

Table 3. Ablation study on different datasets with 0.9 missing rate.

Dataset	EXOTIC-1	EXOTIC-2	EXOTIC-3	EXOTIC-4	EXOTIC
Caltech101	85.47	84.31	86.13	85.38	**88.16**
Scene15	62.35	62.31	62.77	**64.72**	62.53
LandUse21	54.14	51.05	53.38	54.43	**55.24**
HW	93.95	92.80	93.10	93.30	**94.20**
Fashion	91.10	92.03	90.47	90.85	**92.55**
NUSWIDE	**44.12**	42.25	42.40	39.06	43.90
CUB	80.83	80.00	**81.17**	79.33	81.00

helps enhance performance. However, when α and β become too large, the external knowledge incorporates excessive task-irrelevant information, introducing redundancy and suppressing classification performance. When q becomes too large, the excessively sharp penalty leads to a decline in effectiveness.

4.6. Ablation Studies

To evaluate the contribution of each component, ablation experiments are conducted under four variants. Specifically, EXOTIC-1, EXOTIC-2, and EXOTIC-3 denote the variants without knowledge alignment loss $\mathcal{L}_{kc}$, category alignment loss $\mathcal{L}_{vc}$, and both losses, respectively, while EXOTIC-4 denotes removing external knowledge completion. As shown in Tab. 3, we can observe four findings: (1) The removal of the loss term $\mathcal{L}_{kc}$ results in noticeable performance drops across most datasets, indicating its importance in balancing internal and external information and preventing semantic conflicts. (2) When the loss term $\mathcal{L}_{vc}$ is removed, noticeable performance drops occur across all datasets, especially on the LandUse21 dataset, showing its role in enhancing the task relevance of external knowledge. (3) When both losses are removed, accuracy declines further, confirming their complementary effect to enhance classification accuracy. 4) When external knowledge completion is not employed, performance decline demonstrates the powerful role of knowledge completion in handling missing information.

5. Conclusion

In this paper, we propose a novel incomplete multi-view classification method (EXOTIC) that incorporates external vision knowledge guidance to impute missing views. Specifically, EXOTIC includes three key contributions: (1) We introduce a vision-driven paradigm that constructs an external knowledge library from large-scale unlabeled images, enabling rich semantic priors to guide incomplete multi-view learning. (2) We design a Knowledge Filtering and Knowledge Purification module to adaptively filter, purify external knowledge, effectively mitigating semantic inconsistency and conflicts. (3) We propose an External Completion strategy that leverages purified external knowledge to impute missing views, achieving robust and accurate classification. Extensive experiments on multiple benchmarks demonstrate the superior performance and robustness of EXOTIC, particularly under high missing rates.

Acknowledgment This work is supported by the Sichuan Science and Technology Planning Project (Grant No. 2026NSFSC1480 and 2024NSFTD0049), Central Government's Guide to Local Science and Technology Development Fund under Grant 2025ZYDF101, Chengdu Science and Technology Project (Grants No. 2025-YF08-00104-GX and 2025-YF05-00169-SN)

References

[1] Shaotian Cai, Liping Qiu, Xiaojun Chen, Qin Zhang, and Longteng Chen. Semantic-enhanced image clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6869–6878, 2023. 2, 4

[2] Guanqun Cao, Alexandros Iosifidis, Ke Chen, and Moncef Gabbouj. Generalized multi-view embedding for visual recognition and cross-modal retrieval. *IEEE Transactions on Cybernetics*, 48(9):2542–2555, 2017. 2

[3] Xiaochun Cao, Changqing Zhang, Huazhu Fu, Si Liu, and Hua Zhang. Diversity-induced multi-view subspace clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–594, 2015. 2

[4] Guoqing Chao, Yi Jiang, and Dianhui Chu. Incomplete contrastive multi-view clustering with high-confidence guiding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11221–11229, 2024.

[5] Jie Chen, Hua Mao, Dezhong Peng, Changqing Zhang, and Xi Peng. Multiview clustering by consensus spectral rotation fusion. *IEEE Transactions on Image Processing*, 32:5153–5166, 2023. 2

[6] Yanglin Feng, Yongxiang Li, Yuan Sun, Yang Qin, Dezhong Peng, and Peng Hu. Interactive cross-modal learning for text-3d scene retrieval. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 2

[7] Yanglin Feng, Hongyuan Zhu, Dezhong Peng, Xi Peng, Xiaomin Song, and Peng Hu. Robust cross-modal alignment learning for cross-scene spatial reasoning and grounding. In *The Thirtyninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[8] Haijuan Fu, Yu Geng, Changqing Zhang, Zechao Li, and Qinghua Hu. Red-nets: Redistribution networks for multi-view classification. *Information Fusion*, 65:119–127, 2021. 2

[9] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2551–2566, 2022. 2

[10] Changhao He, Hongyuan Zhu, Peng Hu, and Xi Peng. Robust variational contrastive learning for partially view-unaligned clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4167–4176, 2024. 2

[11] Haojian Huang, Chuanyu Qin, Zhe Liu, Kaijing Ma, Jin Chen, Han Fang, Chao Ban, Hao Sun, and Zhongjiang He. Trusted unified feature-neighborhood dynamics for multi-view classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17413–17421, 2025. 2

[12] Zhangqi Jiang, Tingjin Luo, and Xinyan Liang. Deep incomplete multi-view learning network with insufficient label information. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 12919–12927, 2024. 2

[13] Haobin Li, Yunfan Li, Mouxing Yang, Peng Hu, Dezhong Peng, and Xi Peng. Incomplete multi-view clustering via prototype-based imputation. In *Proceedings of the 32th International Joint Conference on Artificial Intelligence*, 2023. 3

[14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 4

[15] Shuxian Li, Changhao He, Xiting Liu, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Learning with noisy triplet correspondence for composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 19628–19637, 2025. 2

[16] Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Jianping Fan, and Xi Peng. Image clustering with external guidance. In *Forty-first International Conference on Machine Learning*, 2024. 2, 4

[17] Ke Liang, Lingyuan Meng, Meng Liu, Yue Liu, Wenxuan Tu, Siwei Wang, Sihang Zhou, and Xinwang Liu. Learn from relational correlations and periodic events for temporal knowledge graph reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1559–1568, 2023. 2

[18] Ke Liang, Lingyuan Meng, Hao Li, Meng Liu, Siwei Wang, Sihang Zhou, Xinwang Liu, and Kunlun He. Mgksite: Multimodal knowledge-driven site selection via intra and inter-modal graph fusion. *IEEE Transactions on Multimedia*, 27: 1722–1735, 2025. 1

[19] Ke Liang, Lingyuan Meng, Hao Li, Jun Wang, Long Lan, Miaomiao Li, Xinwang Liu, and Huaimin Wang. From concrete to abstract: multi-view clustering on relational knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[20] Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4447–4461, 2022. 2, 7

[21] Chengliang Liu, Jie Wen, Xiaoling Luo, Chao Huang, Zhihao Wu, and Yong Xu. Dicnet: Deep instance-level contrastive network for double incomplete multi-view multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8807–8815, 2023. 2, 3, 7

[22] Chengliang Liu, Jie Wen, Xiaoling Luo, and Yong Xu. Incomplete multi-view multi-label learning via label-guided masked view-and category-aware transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8816–8824, 2023. 3, 7

[23] Chengliang Liu, Jinlong Jia, Jie Wen, Yabo Liu, Xiaoling Luo, Chao Huang, and Yong Xu. Attention-induced embedding imputation for incomplete multi-view partial multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13864–13872, 2024. 3, 7

[24] Chengliang Liu, Jie Wen, Yabo Liu, Chao Huang, Zhihao Wu, Xiaoling Luo, and Yong Xu. Masked two-channel decoupling framework for incomplete multi-view weak multi-label learning. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 7

[25] Chengliang Liu, Gehui Xu, Jie Wen, Yabo Liu, Chao Huang, and Yong Xu. Partial multi-view multi-label classification

via semantic invariance learning and prototype modeling. In *Forty-first International Conference on Machine Learning*, 2024. 7

[26] Chengliang Liu, Jie Wen, Yong Xu, Bob Zhang, Liqiang Nie, and Min Zhang. Reliable representation learning for incomplete multi-view missing multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4940 – 4956, 2025. 7

[27] Wei Liu, Yufei Chen, Xiaodong Yue, Changqing Zhang, and Shaorong Xie. Safe multi-view deep classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8870–8878, 2023. 2

[28] Wei Liu, Yufei Chen, and Xiaodong Yue. Enhancing multi-view classification reliability with adaptive rejection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18969–18977, 2025. 2

[29] Yiding Lu, Haobin Li, Yunfan Li, Yijie Lin, and Xi Peng. A survey on deep clustering: from the prior perspective. *Vicinagearth*, 1(1):4, 2024. 2

[30] Chao Su, Likang Peng, Yuan Sun, Dezhong Peng, Xi Peng, and Xu Wang. Neighbor-aware contrastive disambiguation for cross-modal hashing with redundant annotations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 1

[31] Chao Su, Zhi Li, Tianyi Lei, Dezhong Peng, and Xu Wang. Metavg: A meta-learning framework for visual grounding. *IEEE Signal Processing Letters*, 31:236–240, 2024. 1

[32] Chao Su, Huiming Zheng, Dezhong Peng, and Xu Wang. Dica: Disambiguated contrastive alignment for cross-modal retrieval with partial labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 20610–20618, 2025. 2

[33] Chao Su, Yanan Li, Xu Wang, Yingke Chen, Huiming Zheng, Dezhong Peng, and Yuan Sun. Ambiguity-tolerant cross-modal hashing with partial labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. 1

[34] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International Conference on Machine Learning*, pages 1083–1092. PMLR, 2015. 3

[35] Jie Wen, Chengliang Liu, Shijie Deng, Yicheng Liu, Lunke Fei, Ke Yan, and Yong Xu. Deep double incomplete multi-view multi-label learning with incomplete labels and missing views. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 3, 7

[36] Liang Xie, Jialie Shen, Jungong Han, Lei Zhu, and Ling Shao. Dynamic multi-view hashing for online image retrieval. In *IJCAI*, 2017. 2

[37] Mengyao Xie, Zongbo Han, Changqing Zhang, Yichen Bai, and Qinghua Hu. Exploring and exploiting uncertainty for incomplete multi-view classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19873–19882, 2023. 7

[38] Wulin Xie, Lian Zhao, Jiang Long, Xiaohuan Lu, and Bingyan Nie. Multi-view factorizing and disentangling: A novel framework for incomplete multi-view multi-label classification. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1914–1923. IEEE, 2025. 2

[39] Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. Reliable conflictive multi-view learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16129–16137, 2024. 2, 3

[40] Cai Xu, Yilin Zhang, Ziyu Guan, and Wei Zhao. Trusted multi-view learning with label noise. In *IJCAI*, pages 5263–5271, 2024. 3

[41] Deng Xu, Chao Zhang, Cong Guo, Chunlin Chen, and Huaxiong Li. Fast incomplete multi-view clustering with adaptive similarity completion and reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21734–21742, 2025. 2

[42] Shilin Xu, Yuan Sun, Xingfeng Li, Siyuan Duan, Zhenwen Ren, Zheng Liu, and Dezhong Peng. Noisy label calibration for multi-view classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21797–21805, 2025. 3

[43] Honglin Yuan, Shiyun Lai, Xingfeng Li, Jian Dai, Yuan Sun, and Zhenwen Ren. Robust prototype completion for incomplete multi-view clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10402–10411, 2024. 2, 3

[44] Pengxin Zeng, Mouxing Yang, Yiding Lu, Changqing Zhang, Peng Hu, and Xi Peng. Semantic invariant multi-view clustering with fully incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2139–2150, 2023. 2

[45] Changqing Zhang, Ziwei Yu, Qinghua Hu, Pengfei Zhu, Xinwang Liu, and Xiaobo Wang. Latent semantic aware multi-view multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 3

[46] Changqing Zhang, Zongbo Han, Huazhu Fu, Joey Tianyi Zhou, Qinghua Hu, et al. Cpm-nets: Cross partial multi-view networks. *Advances in Neural Information Processing Systems*, 32, 2019. 7

[47] Chao Zhang, Xiuyi Jia, Zechao Li, Chunlin Chen, and Huaxiong Li. Learning cluster-wise anchors for multi-view clustering. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 16696–16704, 2024. 2

[48] Linhao Zhang, Li Jin, Guangluan Xu, Xiaoyu Li, Cai Xu, Kaiwen Wei, Nayu Liu, and Haonan Liu. Camel: Capturing metaphorical alignment with context disentangling for multimodal emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9341–9349, 2024. 3

[49] Sicheng Zhao, Hongxun Yao, Yanhao Zhang, Yasi Wang, and Shaohui Liu. View-based 3d object retrieval via multimodal graph learning. *Signal Processing*, 112:110–118, 2015. 2

[50] Yiwu Zhong, Zi-Yuan Hu, Michael Lyu, and Liwei Wang. Beyond embeddings: The promise of visual table in visual reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6876–6911, 2024. 4