Global-Semantic Alignment Distillation for Partial Multi-view Classification

Xiaoli Wang[1*], Anqi Huang[1*], Yongli Wang[1†], Guanzhou Ke[2], Xiaobin Hong[3†], Jun Liu[4]

[1]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, 210000, China
[2]School of Economics and Management, Beijing Jiaotong University, Beijing, 100080, China
[3]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210023, China
[4]School of Computing and Communications, Lancaster University, Lancaster, United Kingdom
{xiaoliwang, anqihuang, yongliwang}@njust.edu.cn, guanzhouke@bjtu.edu.cn, xiaobinhong@smail.nju.edu.cn, j.liu81@lancaster.ac.uk

Abstract

Partial multi-view classification (PMvC) poses a significant challenge due to the incomplete nature of multi-view data, which complicates effective information fusion and accurate classification. Existing PMvC methods typically rely on heuristic evaluations of view informativeness to achieve global alignment for downstream classification tasks. However, these approaches suffer from two critical issues: **information redundancy** and **semantic misalignment**. The complexity of missing data not only leads to over-reliance on redundant or less informative views but also exacerbates semantic misalignment across views, making it difficult for existing methods to effectively capture and discriminate the task-related features. To address these issues, this work proposes a novel **GL**obal-semantic **A**lignment **D**istillation (GLAD) paradigm for partial multi-view classification, implemented in an imputation-free manner. Our approach incorporates a self-distillation mechanism that enables the model to extract informative features and achieve global semantic alignment across views. The key insight of GLAD is leveraging the ground truth as semantic anchors to guide the alignment of partial multi-view features. By integrating the high-level semantics with extracted features via a cross-attention mechanism, we generate ideal embeddings that consistently capture global semantics across views. These embeddings then serve as intermediate supervision for distilling the student model, ensuring robust semantic alignment even with missing views. Furthermore, we introduce a margin-aware weighting strategy to enhance the model's discriminative ability. Extensive experimental results validate the effectiveness and superiority of the proposed method, showcasing significant improvements in classification performance over existing techniques.

Introduction

Multi-view learning aims to extract a comprehensive understanding of an object by leveraging different views or modalities (Tang et al. 2024; Ke et al. 2023; Xu et al. 2024c; Gu, Li, and Feng 2024; Liu et al. 2023b; Xu et al. 2024b). Due to the diversity of sensor devices and expressions, multi-view learning has attracted considerable attention and demonstrated significant success in various real-world applications, such as clinical diagnostics (Zhou et al. 2023), autonomous driving (Cui et al. 2024), etc. However, most existing methods operate under the assumption of complete views, which often fails to be satisfied in real-world scenarios. In practice, view incompleteness is almost inevitable, arising from issues such as privacy concerns, data corruption, or sensor failures. This presents a challenging research question: *How to efficiently extract and integrate informative features for PMvC?*

Existing PMvC solutions typically fall into two categories: *imputation-based* (Xie et al. 2023a; Liu et al. 2024) and *imputation-free* (Liu et al. 2023a; Wen et al. 2023) methods. *Imputation-based* methods tend to recover the missing data by mining the latent relationships across views or samples, and then utilize imputed complete data for classification (Zhang and Chen 2022; Lin et al. 2022). For example, VIST (Ou et al. 2024) estimates the data distribution from the closest K category vectors of observable views and complements missing views by sampling. While DCP (Lin et al. 2022) recovers the missing views by minimizing the conditional entropy through dual prediction. On the other hand, *imputation-free* methods generally neglect missing views and treat the remaining data under task supervision (Zhu et al. 2022). In this line, CPM-Nets (Zhang et al. 2019) ignores missing views, focusing on common representation learning for downstream tasks. We argue that the existing partial multi-view classification methods suffer from the issues of **information redundancy** and **semantic misalignment**. Although imputation-based approaches effectively manage missing views, they come with inherent drawbacks such as privacy risks, computational overhead, noise perturbation, and task-irrelevant information. These limitations hinder their performance and practical deployment, particularly in areas with safety-critical and privacy-sensitive concerns. On the other hand, while imputation-free methods avoid redundancy pretext tasks, their performance mainly stems from heuristically evaluating the informativeness of different views to perform global alignment for the downstream classification task, which presents a formidable challenge in remaining the low-quality data scenario.

In light of these limitations, this study aims to develop a unified framework that integrates the extraction of informative features and inter-view semantic alignment in PMvC. A key challenge lies in providing additional discriminative guidance, as relying solely on task supervision proves insufficient

[*]These authors contributed equally.
[†]Corresponding authors.

for distinguishing critical information and achieving semantic alignment. This challenge becomes even more pronounced with the absence of relevant modalities, which further complicates the extraction of task-related features. To address these issues, we propose a novel self-distillation method for PMvC, termed **GL**obal-semantic **A**lignment **D**istillation (GLAD), as illustrated in Figure 1. GLAD comprises a two-phase training process followed by an inference step: (1) *Global-semantic alignment teacher learning*: We enhance the extraction and alignment of semantic information by incorporating ground truth labels as discriminative guidance. This approach helps in capturing informative features from partial multi-view embeddings and generates ideal embeddings that represent comprehensive global semantic alignment across views. (2) *Margin-aware distillation-based student learning*: The ideal embeddings serve as intermediate supervision distilling for the student model, ensuring it captures comprehensive semantic information even with partial multi-view data. To improve the inter-class discriminability of the model, we introduce a margin-aware distillation weighting strategy, refining the distillation process and facilitating the generation of class-friendly embeddings. (3) During inference, the student model, trained without label inputs, serves as the reference model, ensuring no data leakage risk. Overall, the contributions of this work are summarized as follows:

- We propose a global-semantic alignment distillation method GLAD tailored for partial multi-view classification without view imputation. Our method addresses the view incompleteness challenges by enabling the precise extraction of task-relevant features, thereby achieving global-semantic alignment across views.
- By leveraging labels as semantic anchors, our method performs global alignment among partial multi-view representations. We further introduce a margin-aware weighting-driven distillation loss to facilitate a deeper understanding of alignment semantics, crucial for accurate classification. GLAD works in a self-distillation framework, enabling no data leakage risk.
- We conduct extensive experiments and the results substantiate the effectiveness and superiority of the proposed method, showcasing enhancements and improvements in classification compared to existing approaches.

Related Works

Partial Multi-view Classification

A precise understanding of data plays a critical role in enhancing the performance of downstream tasks (Hong et al. 2021b,a; Li, Tang, and Mei 2018; Li et al. 2021). Multi-view data offers the advantage of providing complementary information, enabling a more comprehensive understanding (Mo et al. 2023; Pan and Kang 2021; Wang et al. 2024b). However, inevitable missing views result in incomplete views, hindering holistic comprehension. Consequently, the classification of partial multi-view data has become a common challenge in real-world applications. Existing approaches to partial multi-view classification are typically divided into imputation-based and imputation-free methods, based on their treatment of missing views.

Imputation-based PMvC methods recover missing views by mining latent relationships across views or samples (Ou et al. 2024; Zhang and Chen 2022). For example, DCP (Lin et al. 2022) learned view-specific representations and recovered missing views by minimizing conditional entropy between views, enabling mutual prediction in the latent space. However, imputing missing views introduces redundant information, as accurately estimating these views without ground truth is particularly challenging. Furthermore, a high rate of missing data increases computational complexity, which hinders the practical deployment of such methods, especially in security-critical domains. Imputation-free methods, by contrast, avoid redundancy by directly integrating observed views to learn latent representations (Zhu et al. 2022; Zhang et al. 2020). In this line, DICNet (Liu et al. 2023a) learned view-specific representations from observable views, explored inter-view consistency through contrastive learning, and fused these features for classification tasks. While these methods avoid the challenges of imputation, their performance heavily relies on heuristically evaluating the informativeness of different views to achieve global alignment for downstream classification tasks. This poses significant difficulties in remaining low-quality data scenarios.

In contrast to these approaches, our method adopts an imputation-free strategy from the outset, avoiding redundancy associated with imputation. By incorporating labels directly into the input, we provide discriminative guidance that facilitates the learning of informative features from partial multi-view representations. This fine-grained reference allows for global-semantic alignment across views, enhancing the model's ability to focus on discriminative vectors. Consequently, our approach improves classification performance in incomplete data scenarios.

Label Enhancement Methods

In the supervision learning paradigm, labels serve as the explicit cues for formulating the objective function, play a crucial role in facilitating the optimization of model parameters. Recently, numerous studies explore label-enhanced mechanisms to improve model training (Yang et al. 2021; You et al. 2020; Bengio, Weston, and Grangier 2010; Sun et al. 2017; Li et al. 2022b). For example, LabelEnc (Hao et al. 2020) introduces a label encoding function to enhance the training of object detection systems. LAD-GNN (Hong et al. 2024) proposes a label-attentive approach to boost GNN learning for graph-level tasks. CMA (You et al. 2020) constructs a label graph and learns semantic label embeddings to guide cross-modality attention learning, thereby enhancing multi-label classification performance. However, these methods focus on single-modality data, which limits their effectiveness in handling multi-view data. Multi-view data presents complex inter-view relationships and challenges in adequately exploring cross-view consistency and diverse information, especially when some views are missing. In contrast, this work is the first to leverage labels as semantic anchors to enhance the performance of partial multi-view classification, guiding global-semantic alignment across views.

Knowledge Distillation

Knowledge distillation is a technique that transfers knowledge from a complex, high-performing model, known as the teacher model, to a smaller and more computationally efficient model, called the student model. This process reduces the complexity of the model and the computational resources needed while aiming to retain the original performance levels of the model (Xie et al. 2024, 2023b). Knowledge distillation is widely used across various domains, including object detection (Li et al. 2022a, 2023), semantic segmentation (Gao et al. 2024; Li, Halstead, and Mccool 2024), and multi-view learning (Wang et al. 2024a). The first extension of knowledge distillation to multi-view learning challenges comes with MTS-Net (Tian, Sun, and Tang 2022). KDMVC (Wang et al. 2024a) uses self-knowledge distillation within the context of semi-supervised multi-view learning to improve classification performance. While existing methods typically use soft labels produced by the teacher model for distillation, our approach allows the student model to acquire knowledge by emulating the ideal embeddings generated by the teacher model, offering a more efficient and informative reference.

Methodology

Problem Formulation

Given a multi-view training set with N samples, denoted as $\{\mathbf{X}_n, \mathbf{y}_n\}_{n=1}^N$, where $\mathbf{X}_n = \{\mathbf{x}_n^v \in \mathbb{R}^{d_v}\}_{v=1}^V$ consists of V views, d_v is the feature dimension of the samples in the v-th view and $\mathbf{y}_n$ is class label. To indicate whether a view is missing, an indicator matrix $\mathbf{I}$ is introduced, where $\mathbf{I} \in \{0,1\}^{N\times V}$, with $\sum_{j=1}^V \mathbf{I}_{ij} \geq 1$ for $\forall i$. For example: $\mathbf{I}_{ij} = 0$ indicates that the j-th view of the i-th sample is missing, while $\mathbf{I}_{ij} = 1$ indicates that the j-th view of the i-th sample is observed. During the data pre-processing phase, missing views are filled with 0. The PMvC task aims to train a model on incomplete multi-view data that can accurately classify new samples with arbitrary missing view patterns.

Overview

In this section, we introduce GLAD for partial multi-view classification. The overall framework is depicted in Figure 1, where the teacher and student models train iteratively within a uniform self-distillation framework. Specifically, the student network is the inference network. Our method is devoid of imputation techniques to circumvent the limitations associated with view-imputation methods, including the potential for introducing redundant information. GLAD employs the ground-truth to address semantic misalignment and reduce redundant information. The approach comprises two distinct learning phases: (1) Global-semantic alignment teacher learning and (2) Margin-aware distillation-based student learning. The teacher model is designed to leverage label embeddings as discriminative guidance, facilitating global-semantic alignment across different views and producing ideal embeddings. This setup allows the student model to absorb comprehensive semantic information through margin-aware distillation. Ultimately, the student model is refined via both distillation and task supervision, enhancing its ability to differentiate category-specific features and mitigate semantic misalignment.

Global-Semantic Alignment Teacher Learning

This section introduces the teacher model that transfers comprehensive semantic information to the student model. The teacher model comprises two branches: the first branch takes the ground-truth as input and projects them into label embeddings using the label encoder, while the second branch processes partial multi-view data, tokenizing each view as a token for transformer input. The global-semantic alignment component employs cross-attention to fuse label embeddings and extracted partial multi-view embeddings. This fusion process generates ideal embeddings that encapsulate global-semantic alignment across views. These ideal embeddings are then fed into the classification head to produce classification results. Both branches are jointly trained to minimize classification loss.

Multi-view Embeddings. Cross-view interactions enable the model to leverage shared and diverse features from each view. To achieve this, we project multiple views into the same dimensional embeddings, facilitating better alignment and exploring cross-view intrinsic relationships. For each multi-view data $\mathbf{x}_i^v$ of v-th view, view-specific fully connected layers (FC) are employed to project it into view-specific embeddings:

$$\mathbf{z}_i^v = \text{LN}(f^v(\mathbf{x}_i^v)), \tag{1}$$

where LN represents *layer normalization*. $\mathbf{z}^v \in \mathbb{R}^{N\times D}$ denotes the v-th view embeddings with D-dimensional. Therefore, multi-view embeddings can be formulated as: $\mathbf{Z}^t = [\mathbf{z}^1, \mathbf{z}^2, \ldots, \mathbf{z}^V]$, where $\mathbf{Z}^t \in \mathbb{R}^{N\times V\times D}$ and $[\cdot,\cdot]$ denotes the concatenation operation performed along the rows.

Global-Semantic Alignment. Our teacher model introduces groud-truth labels as a form of discriminative guidance and guides the cross-view global-semantic alignment process. Given a ground-truth label $\mathbf{y}_i$, we employ an isolate label encoder, equipped with a Multi-Layer Perceptron (MLP), to map this label into label embeddings:

$$\mathbf{H}_i^{(L)} = E(\mathbf{y}_i), \tag{2}$$

where $\mathbf{H}^{(L)} \in \mathbb{R}^{N\times D}$ represents the label embeddings, and $E(\cdot)$ defines the label encoder. These embeddings $\mathbf{H}^{(L)}$ encapsulate global-semantic information that is crucial for enhancing the discrimination of category-specific features within the multi-view embeddings $\mathbf{Z}^t$. To leverage this, we introduce an attention mechanism to capture the intrinsic relationships between label embeddings and multi-view embeddings. Initially, we apply a set of linear layers with weights $\{\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V\}$ to map the label embeddings $\mathbf{H}_i^{(L)}$ into *queries* and the multi-view embeddings $\mathbf{Z}_i^t$ into *keys* and *values*. The attention mechanism is then formulated as:

$$\begin{aligned}\mathbf{H}_i^t &= \text{Attention}(\mathbf{Z}_i^t, \mathbf{H}_i^{(L)}) \\ &= \text{Softmax}\left(zerofill(\frac{QK^T}{\sqrt{d_k}})\cdot\tau\right)V,\end{aligned} \tag{3}$$

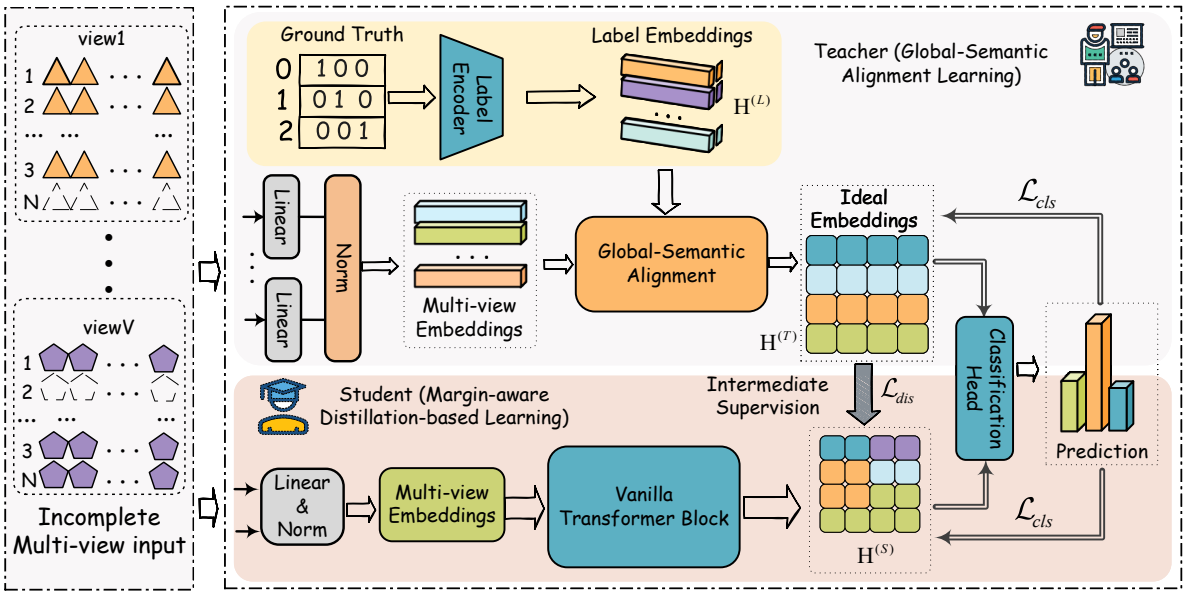

Figure 1: An overview of our GLAD. It comprises two components: the student network, utilized for final inference without label inputs, and the teacher network, responsible for transferring comprehensive semantic information to the student network.

where $\mathbf{H}^t$ represents the fused embeddings. $K = \mathbf{W}_K \mathbf{Z}_i^t$, $V = \mathbf{W}_V \mathbf{Z}_i^t$, and $Q = \mathbf{W}_Q \mathbf{H}_i^{(L)}$. The parameter τ is the attention scaling coefficient (Zhang et al. 2022). The function *zerofill* fills missing views with $-1e9$ to mask them. Subsequently, we enhance the fusion process by applying an *add & layer normalization* (LN) operation followed by a *feed-forward network* (FFN). This process is described as:

$$\mathbf{H}_i^{t'} = \text{FFN}(\text{LN}(\mathbf{H}_i^t + \mathbf{Z}_i^t)) + \mathbf{H}_i^t, \tag{4}$$

To account for missing views, we employ the indicator matrix $\mathbf{I}$ to weighted fusion of these embeddings across views. The final ideal embeddings are computed as:

$$\mathbf{H}_i^{(T)} = \frac{1}{V} \sum_{v=1}^{V} \mathbf{I}_{i,v} \mathbf{H}_{i,v,:}^{t'} \tag{5}$$

where $\mathbf{H}^{(T)} \in \mathbb{R}^{N \times D}$ denotes the ideal embeddings that encapsulate global-semantic alignment across views.

Teacher Model Training. The ideal embeddings $\mathbf{H}_i^{(T)}$ are then fed into a shared classification head to produce the predicted label p_i^t. The teacher model is optimized using cross-entropy, which is formulated as:

$$\mathcal{L}_{cls} = \frac{1}{N} \sum_{i=1}^{N} -\mathbf{y}_i log(p_i^t) \tag{6}$$

Margin-aware Distillation-based Student Learning

Once the teacher model converges, the ideal embeddings $\mathbf{H}^{(T)}$ are used as intermediate supervision to distill the student model, ensuring it captures comprehensive semantic information despite the partial multi-view data. To refine the knowledge transfer process, we employ a margin-aware distillation strategy. The core idea is to encourage the model to focus more on samples that are near the decision boundaries. Additionally, the student model shares the same classification head with the teacher model. During student model training, both the distillation loss and classification loss are jointly optimized.

Margin-aware Distillation. Given the multi-view data $\{\mathbf{x}^v\}_{v=1}^V$, we first project each view into view-specific embeddings and concatenate these to obtain the multi-view embeddings $\mathbf{Z}^{\mathbf{s}} \in \mathbb{R}^{N \times V \times D}$, similar to the teacher model. Unlike the teacher model, the student model does not receive label inputs during inference. We utilize self-attention to facilitate cross-view interactions, which helps reduce redundant information and enhances the mining of consistent and diverse cross-view information. The architecture of the student model is as same as that of the teacher model, with the key difference being that in the attention computation of the student model, the multi-view embeddings serve simultaneously as *query*, *key*, and *value*. As a result, we obtain the fused multi-view embeddings $\mathbf{H}^{(S)} \in \mathbb{R}^{N \times D}$.

Given the inherent difficulty in estimating irregular decision boundaries, we draw inspiration from previous work (Litrico, Del Bue, and Morerio 2023; Wei, Luo, and Luo 2023) and introduce the classification uncertainty of each sample from the teacher model as a metric to re-weight the distillation loss. The intuition here is that samples near the decision boundary exhibit higher classification uncertainty. The distillation loss is therefore formulated using Mean Square Error (MSE) as follows:

$$\mathcal{L}_{dis} = \frac{1}{N} \sum_{i=1}^{N} \frac{\mathcal{H}(p_i^t)}{log_2 C} \|\mathbf{H}_i^{(T)}, \mathbf{H}_i^{(S)}\|_2^2, \tag{7}$$

where $\mathcal{H}(\cdot)$ denotes the information entropy and C is the number of the classes in the dataset. This refined distillation loss encourages the student model to improve its inter-class discriminability.

Student Model Training In the student model, both the distillation loss $\mathcal{L}_{dis}$ and the classification loss $\mathcal{L}_{cls}$ are jointly optimized. The distillation loss enables the model to learn comprehensive semantic information from the teacher, while the classification loss drives the model to excel in the partial multi-view classification task. The overall objective function for training the student model is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \cdot \mathcal{L}_{dis}, \tag{8}$$

where $\mathcal{L}_{cls}$ is the classification loss defined in Eq. 6, and λ is the trade-off factor for balancing the two losses, defaulting to 1.

Experiments

Experimental Settings

Datasets. We evaluate the performance of our method on five datasets. **Scene15** (Fei-Fei and Perona 2005): A scene dataset comprising 3 views, containing 15 classes and a total of 4,485 samples. **Animal** (Lampert, Nickisch, and Harmeling 2013): The dataset with 2-views, containing 50 classes and a total of 10,158 samples. **Caltech101** (Fei-Fei, Fergus, and Perona 2004): A subset of the Caltech101 dataset, containing 2,386 samples with 6 views per sample across 20 classes. **BDGP** (Cai et al. 2012): A dataset consisting of images related to Drosophila embryos, containing 2,500 samples across 5 categories. Each sample is composed of 4 views. **LandUse21** (Yang and Newsam 2010): A satellite image dataset comprising 3 views, containing 21 classes with a total of 2,100 samples.

Compared Methods. To demonstrate the effectiveness and superiority of the proposed method, we compare it with one baseline method and seven existing state-of-the-art methods. i.e., (1) **Mean-Imputation** that imputes missing views with the mean of all observed samples of the i-th view; (2) **CPM-Nets** (Zhang et al. 2019) is an imputation-free method that learns latent representations for all views with available data and maps these latent representations to classification predictions; (3) **TMC** (Han et al. 2020) is a decision fusion method that accurately identifies and fuse confident views while ignoring unreliable views; (4) **Mmydynamics** (Han et al. 2022) dynamically evaluates informativeness for each sample and applies weighted fusion multiple views; (5) **DD-IMvMLC-net** (Wen et al. 2023) is an imputation-free method that applies weighted fusion of available views and ignores missing views; (6) **DICNet** (Liu et al. 2023a) aims to enhance the consistency of view-specific features extracted from observable views. Subsequently, these features are concatenated to facilitate classification tasks; (7) **UIMC** (Xie et al. 2023a) samples multiple times from the estimated distribution of missing views to impute them and introduces an evidence-based fusion strategy to integrate multiple views reliably; (8) **RCML** (Xu et al. 2024a) is a decision fusion method that provides decision results and reliabilities despite conflictive multi-view data.

Implementation Details. *Partial Multi-view Data Construction*: To construct partial multi-view datasets, we follow the approach described in (Xie et al. 2023a). Specifically, we randomly select a portion of the samples from a given dataset and remove some of the views from those samples, ensuring that each sample retains at least one view. For a dataset with N samples and V views, the missing rate η is calculated as: $\eta = \frac{\sum_{i=1}^{V} M_i}{V \times N}$, where M_i represents the number of missing samples in i-th view. In the *performance comparison* experiment, we utilize classification accuracy (ACC) as the evaluation metric, following prior works (Xie et al. 2023a). For the imputation-based method UIMC, we follow the settings of original papers to impute the missing views. The imputation-free methods, such as CPM-Nets, DD-IMvMLC-net, and DICNet are trained directly on incomplete data. For multi-view classification methods, such as TMC, Mmydynamics, and RCML, we apply the Mean-Imputation strategy to fill in the missing views. Each experiment is repeated five times, and the results are averaged and recorded. Our proposed GLAD model is implemented using PyTorch 2.0.1. All experiments are conducted on a PC equipped with an NVIDIA GeForce RTX 3090 GPU.

Experimental Results and Analysis

Performance Comparison. We compare our method with one baseline method and seven competitive methods across five datasets with various missing rates to evaluate performance. The experimental results are summarized in Table 1, where the best values are highlighted in bold. From these results, several key observations can be made: (1) Our method consistently outperforms the comparison methods across all datasets, even in the presence of missing views. For instance, when the missing rate $\eta = 0$, our method achieves a 3.21% improvement in accuracy on the Scene15 dataset and a 5.38% improvement on the LandUse21 dataset compared to the second-best method. Additionally, when the missing rate $\eta = 0.1$, our method shows a 5.10% higher accuracy on the LandUse21 dataset, and when $\eta = 0.5$, it achieves a 2.92% higher accuracy on the Scene15 dataset than the second-best method. These results demonstrate the effectiveness, superiority, and robustness of our method in handling incomplete multi-view data. (2) Figure 2 visualizes the classification accuracy on three different datasets as the missing rate increases. The results clearly show that our method consistently outperforms the comparison methods at all missing rates. This consistent performance can be attributed to our ability to achieve global semantic alignment across views by incorporating labels as discriminative guides. This alignment is particularly effective in overcoming the challenges of extracting inter-class discriminative features with missing views.

Ablation Study To demonstrate the efficacy of the proposed global-semantic alignment distillation strategy, we conduct ablation experiments on four datasets, comparing our model (denoted as "w/(std)") to two baseline settings: (1) a model with global-semantic alignment distillation but without the margin-aware weighting strategy (denoted as "w/(basic)"); (2) a model containing only the student network,

Missing Rates	Method	Scene15	Animal	BDGP	LandUse_21	Caltech101-20
$\eta = 0$	Mean-Imputation	76.14(0.00)	86.9(0.00)	95.2(0.00)	66.19(0.00)	93.29(0.00)
	CPM-Nets (2019)	69.90(0.02)	84.41(0.81)	95.96(1.82)	50.71(2.18)	86.71(6.36)
	TMC (2020)	73.85(1.37)	85.11(0.13)	98.00(0.18)	51.24(3.90)	91.24(1.27)
	Mmdynamics (2022)	77.26(0.00)	86.37(0.00)	98.36(0.00)	76.43(0.00)	95.31(0.00)
	DD-IMvMLC-net (2023)	78.86(0.62)	86.01(0.38)	98.52(0.27)	74.62(1.63)	93.58(0.28)
	DICNet (2023a)	80.29(0.37)	84.13(0.44)	98.44(0.41)	78.00(0.76)	92.33(0.79)
	UIMC (2023a)	77.70(0.00)	OOM	95.40(0.13)	60.19(1.03)	94.97(0.00)
	RCML (2024a)	74.02(0.31)	83.53(0.07)	99.40(0.00)	54.48(0.28)	93.67(0.16)
	Ours	**83.50(0.47)**	**88.54(0.36)**	**99.92(0.18)**	**83.38(0.49)**	**95.77(0.40)**
	$\triangle\%$	3.21	1.64	0.52	5.38	0.46
$\eta = 0.1$	Mean-Imputation	72.58(0.00)	81.44(0.00)	91.80(0.00)	60.00(0.00)	90.78(0.00)
	CPM-Nets (2019)	65.66(0.02)	79.96(1.11)	92.04(3.92)	49.57(2.77)	86.71(1.49)
	TMC (2020)	70.59(1.12)	81.49(0.06)	96.84(0.23)	50.76(3.49)	90.31(1.53)
	Mmdynamics (2022)	74.80(0.00)	82.45(0.00)	98.55(0.00)	71.19(0.00)	94.38(0.00)
	DD-IMvMLC-net (2023)	75.41(0.84)	81.93(0.47)	97.44(0.32)	70.76(2.41)	93.58(0.41)
	DICNet (2023a)	77.57(0.44)	80.15(0.18)	97.72(0.27)	73.57(1.76)	92.41(0.36)
	UIMC (2023a)	75.81(0.01)	OOM	88.40(0.49)	47.62(0.50)	87.59(0.43)
	RCML (2024a)	72.24(0.10)	80.23(0.34)	97.04(0.08)	54.00(0.82)	94.05(0.21)
	Ours	**80.07(0.29)**	**83.79(0.27)**	**98.68(0.18)**	**78.67(1.47)**	**95.14(0.18)**
	$\triangle\%$	2.5	1.86	0.13	5.1	0.76
$\eta = 0.5$	Mean-Imputation	57.19(0.00)	69.03(0.00)	77.8(0.00)	36.43(0.00)	80.29(0.00)
	CPM-Nets (2019)	57.08(0.01)	63.15(1.31)	76.20(1.17)	31.00(1.99)	77.76(3.74)
	TMC (2020)	59.71(1.64)	66.80(0.06)	81.60(0.13)	33.33(1.73)	85.24(0.41)
	Mmdynamics (2022)	63.77(0.00)	68.17(0.00)	81.95(0.00)	51.93(0.00)	89.57(0.00)
	DD-IMvMLC-net (2023)	62.45(1.11)	66.93(0.67)	81.72(0.50)	51.38(1.13)	87.38(0.50)
	DICNet (2023a)	63.30(1.54)	64.37(0.44)	82.68(0.75)	55.05(2.34)	87.55(1.21)
	UIMC (2023a)	62.54(0.02)	OOM	63.44(0.93)	33.05(0.35)	70.31(0.49)
	RCML (2024a)	61.76(0.19)	67.00(0.22)	82.60(0.00)	35.76(0.84)	86.54(0.58)
	Ours	**66.69(0.53)**	**70.27(0.27)**	**83.80(0.35)**	**59.62(0.96)**	**91.19(0.26)**
	$\triangle\%$	2.92	1.24	1.12	4.57	1.62

Table 1: The classification accuracy (mean $\pm$ std) of our method and the compared methods at different missing rates, with the best results highlighted in bold. η represents the missing rate, and OOM indicates out-of-memory.

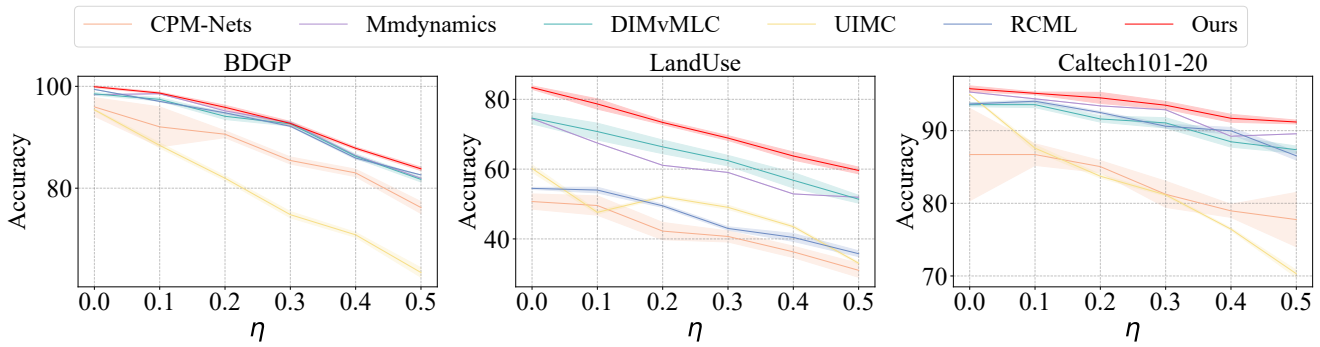

Figure 2: Classification accuracy on BDGP, LandUse21 and Caltech101-20 datasets with different missing rates.

without global-semantic alignment distillation (denoted as "w/o"). The experiments are conducted on complete multi-view samples, where the missing rate $\eta = 0$, and the results are evaluated using ACC. The experimental results are summarized in Table 2. By comparing the results of "w/o" and "w/(basic)", it is evident that the global-semantic alignment distillation strategy significantly improves classification performance. For example, on the LandUse21 and Caltech101-20 datasets, performance improves by 0.71% and 0.45%, respectively. This improvement can be attributed to the naive model captures cross-view consistency information while ignoring important complementary information. In contrast,

our model leverages ideal embeddings as intermediate supervision distilling for the student model, ensuring it captures comprehensive semantic information across views. Additionally, when comparing the results of "w/(basic)" and "w/(std)", it is clear that the margin-aware weighting strategy further enhances classification performance, demonstrating its effectiveness. This indicates that the margin-aware weighting strategy indeed strengthens the model in learning inter-class discriminative representations.

Dataset	w/o	w/(basic)	w/(std)
Scene15	83.39	83.84	83.95
Animal	88.04	88.53	89.12
LandUse21	82.62	83.33	84.52
Caltech101-20	94.97	95.18	96.44

Table 2: Ablation studies on four datasets at $\eta = 0$. "w/(std)" represents our model, "w/(basic)" excludes the margin-aware weighting strategy, and "w/o" uses only the student network.

Visualization To intuitively assess the quality of the partial multi-view representations learned by our method, we visualize the raw view data and the partial multi-view representations learned by our method, as well as by the competitive methods DD-IMvMLC-net and RCML, on the test sets of the BDGP dataset, with a missing rate of $\eta = 0.5$. The visualizations are presented in Figure 3, where the partial multi-view representations are projected into a two-dimensional space using t-SNE. Results illustrate that the partial multi-view embeddings learned by our method exhibit a clearer and more distinct classification structure than the baseline methods. This highlights the ability of our method to learn class-friendly embeddings despite missing views. In contrast, the baseline methods, lacking discriminative guidance, struggle to capture inter-class discriminative features, leading to learned embeddings with a more ambiguous classification structure. The results suggest that utilizing the ground-truth as discriminative guidance significantly enhances the ability of the model to capture inter-class discriminative features. which is crucial for learning robust and class-distinctive representations, especially in scenarios with missing views.

Hyper-parameter Analysis We further discuss the sensitivity of the hyper-parameters λ and τ in Eq. 8 and Eq. 3, with the experimental results recorded in Figure 4. We varied λ from 0.01 to 100 and τ from 0.1 to 0.9, testing the classification performance on three datasets with $\eta = 0$. From the experimental results, we observe that different values of these hyper-parameters affect classification performance differently across datasets. However, the optimal λ can be obtained through tuning on a validation set. The τ values that achieve the best performance vary across datasets, but after carefully considering the results from multiple datasets, we set τ to 0.7 as the standard value for our experiments.

Conclusion

In this work, we propose a novel imputation-free method for the PMvC problem, termed GLobal-semantic Alignment Distillation (GLAD) for partial multi-view classification, which can mitigate semantic misalignment exacerbated by missing views. Specifically, GLAD introduces a self-distillation framework that endows the student model to distinguish informative information and align semantics across views. In the self-distillation framework, this work introduces ground truth labels as discriminative guidance, guiding the generation of ideal embeddings that encapsulate global-semantic alignment across views. These ideal embeddings serve as intermediate supervision distilling for the student model, ensuring it captures comprehensive semantic information. This distillation mechanism enables the student model to generate class-friendly embeddings that significantly improve classification performance. Extensive experimental results validate the effectiveness and superiority of the proposed method.

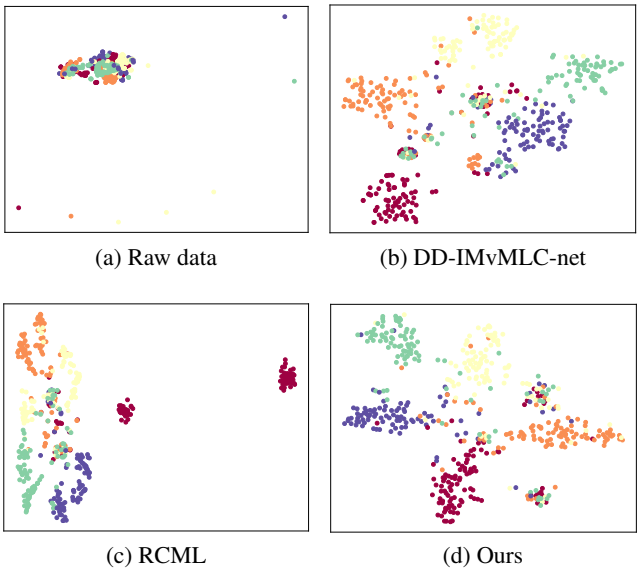

Figure 3: Visualization of concatenated raw data and latent representations learned by different methods on the BDGP dataset with $\eta = 0.5$.

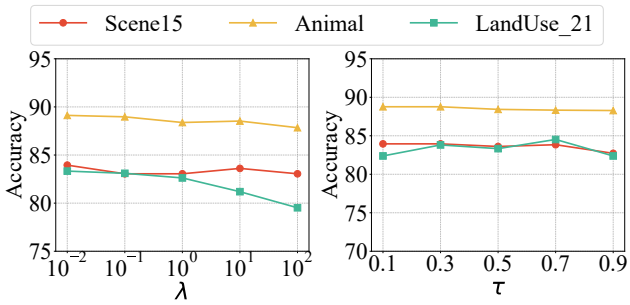

Figure 4: Hyper-parameter sensitivity results of λ and τ on three datasets at $\eta = 0$.

Acknowledgments

This article was partially supported by the Jiangsu Province Key R& D Program (Modern Agriculture) Key Project (BE2023352), Key Medical Research Projects of Jiangsu Provincial Health Commission (ZD2022068), National Natural Science Foundation of China (61941113), the China Scholarship Council under Grant 202306840098. We thank all anonymous reviewers for their constructive comments.

References

Bengio, S.; Weston, J.; and Grangier, D. 2010. Label embedding trees for large multi-class tasks. *Advances in neural information processing systems*, 23.

Cai, X.; Wang, H.; Huang, H.; and Ding, C. 2012. Joint stage recognition and anatomical annotation of drosophila gene expression patterns. *Bioinformatics*, 28(12): i16–i24.

Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Zhou, Y.; Liang, K.; Chen, J.; Lu, J.; Yang, Z.; Liao, K.-D.; et al. 2024. A survey on multimodal large language models for autonomous driving. In *CVPR*, 958–979.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR workshop*, 178–178.

Fei-Fei, L.; and Perona, P. 2005. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, volume 2, 524–531.

Gao, T.; Ao, W.; Wang, X.-A.; Zhao, Y.; Ma, P.; Xie, M.; Fu, H.; Ren, J.; and Gao, Z. 2024. Enrich Distill and Fuse: Generalized Few-Shot Semantic Segmentation in Remote Sensing Leveraging Foundation Model's Assistance. In *CVPR*, 2771–2780.

Gu, Z.; Li, Z.; and Feng, S. 2024. EDISON: Enhanced Dictionary-Induced Tensorized Incomplete Multi-View Clustering with Gaussian Error Rank Minimization. In *Forty-first International Conference on Machine Learning*, 1–9.

Han, Z.; Yang, F.; Huang, J.; Zhang, C.; and Yao, J. 2022. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *CVPR*, 20707–20717.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2020. Trusted multi-view classification. In *ICLR*.

Hao, M.; Liu, Y.; Zhang, X.; and Sun, J. 2020. Labelenc: A new intermediate supervision method for object detection. In *ECCV*.

Hong, X.; Li, W.; Wang, C.; Lin, M.; and Lu, S. 2024. Label Attentive Distillation for GNN-Based Graph Classification. In *AAAI*, volume 38, 8499–8507.

Hong, X.; Zhang, T.; Cui, Z.; Huang, Y.; Shen, P.; Li, S.; and Yang, J. 2021a. Graph game embedding. In *AAAI*, volume 35, 7711–7720.

Hong, X.; Zhang, T.; Cui, Z.; and Yang, J. 2021b. Variational gridded graph convolution network for node classification. *IEEE/CAA Journal of Automatica Sinica*, 8(10): 1697–1708.

Ke, G.; Chao, G.; Wang, X.; Xu, C.; Zhu, Y.; and Yu, Y. 2023. A clustering-guided contrastive fusion for multi-view representation learning. *IEEE Transactions on Circuits and Systems for Video Technology*.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2013. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and machine intelligence*, 36(3): 453–465.

Li, G.; Li, X.; Wang, Y.; Zhang, S.; Wu, Y.; and Liang, D. 2022a. Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation. In *AAAI*, 1306–1313.

Li, M.; Halstead, M.; and Mccool, C. 2024. Knowledge Distillation for Efficient Instance Semantic Segmentation with Transformers. In *CVPR*, 5432–5439.

Li, W.; Chen, J.; Gao, P.; and Huang, Z. 2022b. Label enhancement with label-specific feature learning. *International Journal of Machine Learning and Cybernetics*, 13(10): 2857–2867.

Li, Z.; Sun, Y.; Zhang, L.; and Tang, J. 2021. CTNet: Context-based tandem network for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9904–9917.

Li, Z.; Tang, J.; and Mei, T. 2018. Deep collaborative embedding for social image understanding. *IEEE Transactions on pattern analysis and machine intelligence*, 41(9): 2070–2083.

Li, Z.; Xu, P.; Chang, X.; Yang, L.; Zhang, Y.; Yao, L.; and Chen, X. 2023. When object detection meets knowledge distillation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 10555–10579.

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

Litrico, M.; Del Bue, A.; and Morerio, P. 2023. Guiding Pseudo-labels with Uncertainty Estimation for Source-free Unsupervised Domain Adaptation. In *CVPR*.

Liu, C.; Jia, J.; Wen, J.; Liu, Y.; Luo, X.; Huang, C.; and Xu, Y. 2024. Attention-Induced Embedding Imputation for Incomplete Multi-View Partial Multi-Label Classification. In *AAAI*, volume 38, 13864–13872.

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

Liu, C.; Wen, J.; Wu, Z.; Luo, X.; Huang, C.; and Xu, Y. 2023b. Information recovery-driven deep incomplete multiview clustering network. *IEEE Transactions on Neural Networks and Learning Systems*.

Mo, Y.; Lei, Y.; Shen, J.; Shi, X.; Shen, H. T.; and Zhu, X. 2023. Disentangled multiplex graph representation learning. In *ICML*, 24983–25005. PMLR.

Ou, S.; Xue, Z.; Li, Y.; Liang, M.; Cai, Y.; and Wu, J. 2024. View-Category Interactive Sharing Transformer for Incomplete Multi-View Multi-Label Learning. In *CVPR*, 27467–27476.

Pan, E.; and Kang, Z. 2021. Multi-view contrastive graph clustering. *NeurIPS*, 34: 2148–2159.

Sun, X.; Wei, B.; Ren, X.; and Ma, S. 2017. Label embedding network: Learning label representation for soft training of deep networks. *arXiv preprint arXiv:1710.10393*.

Tang, W.; Li, L.; Liu, X.; Jin, L.; Tang, J.; and Li, Z. 2024. Context disentangling and prototype inheriting for robust visual grounding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(5): 3213–3229.

Tian, Y.; Sun, S.; and Tang, J. 2022. Multi-view teacher-student network. *Neural Networks*, 146: 69–84.

Wang, X.; Wang, Y.; Ke, G.; Wang, Y.; and Hong, X. 2024a. Knowledge distillation-driven semi-supervised multi-view classification. *Information Fusion*, 103: 102098.

Wang, X.; Wang, Y.; Wang, Y.; Huang, A.; and Liu, J. 2024b. Trusted Semi-Supervised Multi-View Classification With Contrastive Learning. *IEEE Transactions on Multimedia*, 26: 8268–8278.

Wei, S.; Luo, C.; and Luo, Y. 2023. MMANet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *CVPR*, 20039–20049.

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*.

Xie, M.; Han, Z.; Zhang, C.; Bai, Y.; and Hu, Q. 2023a. Exploring and exploiting uncertainty for incomplete multi-view classification. In *CVPR*, 19873–19882.

Xie, Y.; Lin, Y.; Cai, W.; Xu, X.; Zhang, H.; Du, Y.; and He, S. 2024. D3still: Decoupled Differential Distillation for Asymmetric Image Retrieval. In *CVPR*, 17181–17190.

Xie, Y.; Zhang, H.; Xu, X.; Zhu, J.; and He, S. 2023b. Towards a Smaller Student: Capacity Dynamic Distillation for Efficient Image Retrieval. In *CVPR*, 16006–16015.

Xu, C.; Si, J.; Guan, Z.; Zhao, W.; Wu, Y.; and Gao, X. 2024a. Reliable conflictive multi-view learning. In *AAAI*, volume 38, 16129–16137.

Xu, J.; Chen, S.; Ren, Y.; Shi, X.; Shen, H.; Niu, G.; and Zhu, X. 2024b. Self-weighted contrastive learning among multiple views for mitigating representation degeneration. *NeurIPS*, 36.

Xu, J.; Ren, Y.; Wang, X.; Feng, L.; Zhang, Z.; Niu, G.; and Zhu, X. 2024c. Investigating and Mitigating the Side Effects of Noisy Views for Self-Supervised Clustering Algorithms in Practical Multi-View Scenarios. In *CVPR*, 22957–22966.

Yang, X.; Hou, L.; Zhou, Y.; Wang, W.; and Yan, J. 2021. Dense label encoding for boundary discontinuity free rotation detection. In *CVPR*, 15819–15829.

Yang, Y.; and Newsam, S. 2010. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 270–279.

You, R.; Guo, Z.; Cui, L.; Long, X.; Bao, Y.; and Wen, S. 2020. Cross-modality attention with semantic graph embedding for multi-label classification. In *AAAI*, 12709–12716.

Zhang, C.; Cui, Y.; Han, Z.; Zhou, J. T.; Fu, H.; and Hu, Q. 2020. Deep partial multi-view learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(5): 2402–2415.

Zhang, C.; Han, Z.; Fu, H.; Zhou, J. T.; Hu, Q.; et al. 2019. CPM-Nets: Cross partial multi-view networks. *NeurIPS*, 32.

Zhang, H.; and Chen, X. 2022. Adaptive incomplete multi-view learning via tensor graph completion. *arXiv preprint arXiv:2208.03710*.

Zhang, S.; Zhang, X.; Bao, H.; and Wei, F. 2022. Attention Temperature Matters in Abstractive Summarization Distillation. In *Proceedings of the Association for Computational Linguistics*, 127–141.

Zhou, H.-Y.; Yu, Y.; Wang, C.; Zhang, S.; Gao, Y.; Pan, J.; Shao, J.; Lu, G.; Zhang, K.; and Li, W. 2023. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nature biomedical engineering*, 7(6): 743–755.

Zhu, P.; Yao, X.; Wang, Y.; Cao, M.; Hui, B.; Zhao, S.; and Hu, Q. 2022. Latent heterogeneous graph network for incomplete multi-view learning. *IEEE Transactions on Multimedia*, 25: 3033–3045.