View-Association-Guided Dynamic Multi-View Classification

Xinyan Liang[1] , **Li Lv**[1] , **Qian Guo**[2*] , **Bingbing Jiang**[3] , **Feijiang Li**[1] , **Liang Du**[1] and **Lu Chen**[1]
[1]Institute of Big Data Science and Industry, Key Laboratory of Evolutionary Science Intelligence of Shanxi Province, Shanxi University, Taiyuan 030006, China
[2] Shanxi Key Laboratory of Big Data Analysis and Parallel Computing, School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China
[3] School of Information Science and Technology, Hangzhou Normal University, Hangzhou, China
{liangxinyan48, lvli924, czguoqian}@163.com, jiangbb@hznu.edu.cn, {fjli, duliang, chenlu}@sxu.edu.cn

Abstract

In multi-view classification tasks, integrating information from multiple views effectively is crucial for improving model performance. However, most existing methods fail to fully leverage the complex relationships between views, often treating them independently or using static fusion strategies. In this paper, we propose a View-Association-Guided Dynamic Multi-View Classification method (AssoDMVC) to address these limitations. Our approach dynamically models and incorporates the relationships between different views during the classification process. Specifically, we introduce a view-association-guided mechanism that captures the dependencies and interactions between views, allowing for more flexible and adaptive feature fusion. This dynamic fusion strategy ensures that each view contributes optimally based on its contextual relevance and the inter-view relationships. Extensive experiments on multiple benchmark datasets demonstrate that our method outperforms traditional multi-view classification techniques, offering a more robust and efficient solution for tasks involving complex multi-view data.

1 Introduction

Rapid advancement in technology has led to an exponential increase in the generation of multi-view data, including images, text, audio, and video [Fu *et al.*, 2025; Jiang *et al.*, 2021; Li *et al.*, 2023; Wen *et al.*, 2023; Zhang *et al.*, 2024a]. This abundance of data presents opportunities and challenges for researchers and practitioners in diverse fields, such as computer vision and natural language processing. Integrating multiple views information from can significantly enhance the performance of machine learning models, improving their ability to understand complex, real-world scenarios [Jiang *et al.*, 2024; Liang *et al.*, 2024; Guo *et al.*, 2024; Fu *et al.*, 2024b].

Most of multi-view classification algorithms exploits two fundamental principles which ensures their success: (i) consensus, and (ii) diversity principles. The consensus principle seeks to maximize the agreement between multiple representations of the data. The diversity principle demonstrates that in a multiview learning problem, each representation or view of the data may contain some information which other views do not have. Based on these two principles, many fusion algorithms for multi-view classification have been proposed. For example, [Liang *et al.*, 2022] introduced the association information between modality features into multi-modal data fusion and proposed an association-based fusion strategy for multi-modal classification (MMC) in an interpretable manner. [Chen *et al.*, 2023] proposed a joint deep learning framework to learn an underlying feature representation from heterogeneous views.

We analyze fusion methods within kernel-based, graphical, encoder-decoder, and attention-based fusion frameworks. Fig. 1 illustrates three typical structures of fusion models [Li and Tang, 2024]. In Fig. 1(a), text and images are processed through simple operations (dot product, multiplication, and addition). In Fig. 1(b), the Fusion Network is designed to combine the individual image and text embeddings. In Fig. 1(c), models use an integrated encoding-decoding process to handle multi-view inputs simultaneously. Most existing fusion methods *implicitly* exploit the relationships between views during the fusion process, where the interactions among views are typically captured without explicit modeling. This implicit approach, while effective in many cases, may not fully capture the nuanced relationships between the different views, leading to suboptimal fusion performance in certain complex multi-view scenarios. The lack of explicit modeling of view relationships could potentially limit the ability of the model to explore consensus or diversity more effectively. To address this limitation, it is crucial to consider explicitly modeling the relationships between views in multi-view fusion frameworks. As shown in Fig. 1(d), we propose an explicit modeling step for the relationships between views before the fusion of the view features.

Moreover, traditional fusion methods have largely overlooked the dynamic variation in the quality of multi-view data. The oversight often results in fusion methods that fail to adapt to the evolving nature of multi-view data, leading to performance degradation in real-world applications. There-

*Corresponding author.

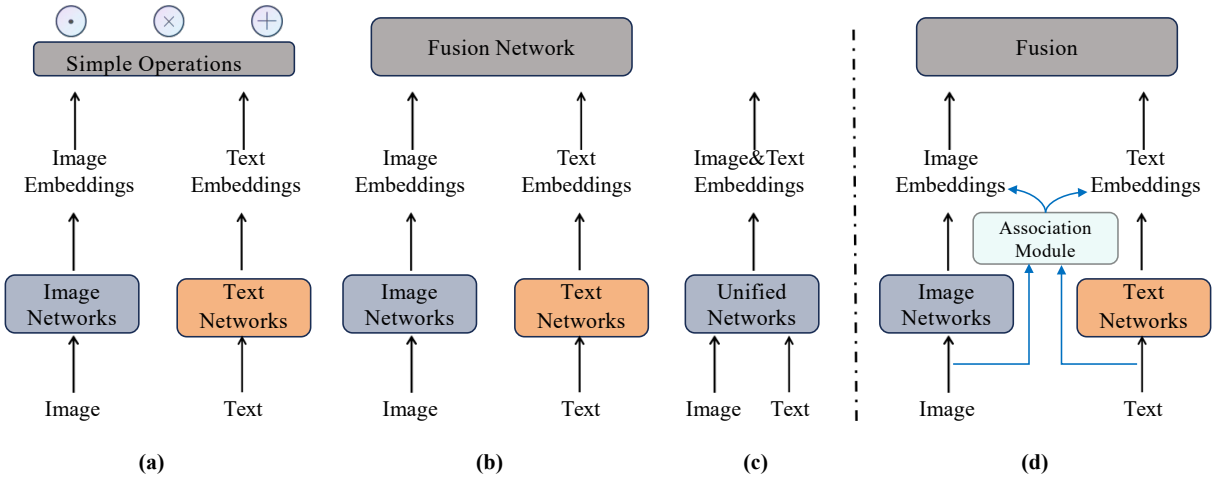

Figure 1: Difference among existing multi-view fusion strategies and ours.

fore, there is a need for fusion techniques that not only integrate information across views but also dynamically adjust to the varying quality of each view.

To address these issues, we propose an end-to-end framework called a view-association-guided dynamic multi-view classification method (AssoDMVC). AssoDMVC is comprised of two fundamental components: a view association encoding module, and a view association-guided dynamic weighting fusion module.

Our main contributions are summarized as follows:

- A view association encoding module is designed to explicitly model the relationships between views, offering guidance for the interaction between different views in downstream tasks, thereby enhancing the effectiveness of the fusion process.
- A novel end-to-end framework, the view-association-guided dynamic multi-view classification method (AssoDMVC), is proposed to address the challenges of multi-view data fusion, providing a more effective and comprehensive solution for integrating diverse views.
- The extensive comparison experiments on six public datasets show that AssoDMVC achieves competitive performance compared to the the state-of-the-art MVC methods.

2 Related Work

Multi-view classification (MVC)is a powerful approach that leverages multiple sources of information (views) to improve the accuracy and robustness of classification tasks. In multi-view classification, each view provides a different perspective or representation of the same underlying data, and combining these views effectively can lead to better performance than relying on a single view[Fu *et al.*, 2024a].

In recent years, many multi-view classification models with complex objectives were further proposed. [Chen *et al.*, 2021] proposed a joint framework for multi-view spectral clustering by learning an adaptive transition probability matrix. The nuclearnorm-based optimization method was proposed to conduct multi-view image data fusion via a joint learning framework[Huang *et al.*, 2020].Sparsity-based optimization methods are also essential in multi-view classification[Wang *et al.*, 2022a].

Fusion methods play a important role in MVC. The quality among multi-view features is different, lots of works consider their contribution to the final tasks[Liang *et al.*, 2025]. Multi-view classification methods can be roughly divided into the feature level fusion-based and decision level fusion-based. Based on where the view contribution are considered, these methods can be grouped into $feature$ level weighting-based method (FW) and $decision$ level weighting-based method (DW). FW learns the contribution weight of each feature[Jiang *et al.*, 2022]. For example, EmbraceNet [Choi and Lee, 2019] assigns 1 to the weight value of one view while 0 to others for each example according to a multinomial distribution. An adaptive-weighting discriminative regression approach (AWDR) [Yang *et al.*, 2019] adopts the square root form of view weight to distinguish features from different views. [Zhang *et al.*, 2024b] proposed discriminative multi-view fusion via adaptive regression (DMVF), it simultaneously discriminates the contribution diversity of different views and samples in an adaptive weighting manner, reducing the influence of low-quality views and outliers for classification. DW learns to assign weights at the decision level. For example, [Han *et al.*, 2022] proposed a trusted multi-view classification (TMC), which models the confidence of each view at an evidence level using the Dempster-Shafer theory.

3 The Proposed Method

In this section, we first introduce a view association encoding module, followed by the design of a view association-guided dynamic weighting fusion module. Fig. 2 provides an overall illustration of the proposed approach.

3.1 Basic Setting

The remainder of this section uses the following notation. Let $\mathcal{X} = \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \times \cdots \times \mathbb{R}^{m_{|V|}}$ represent the instance space

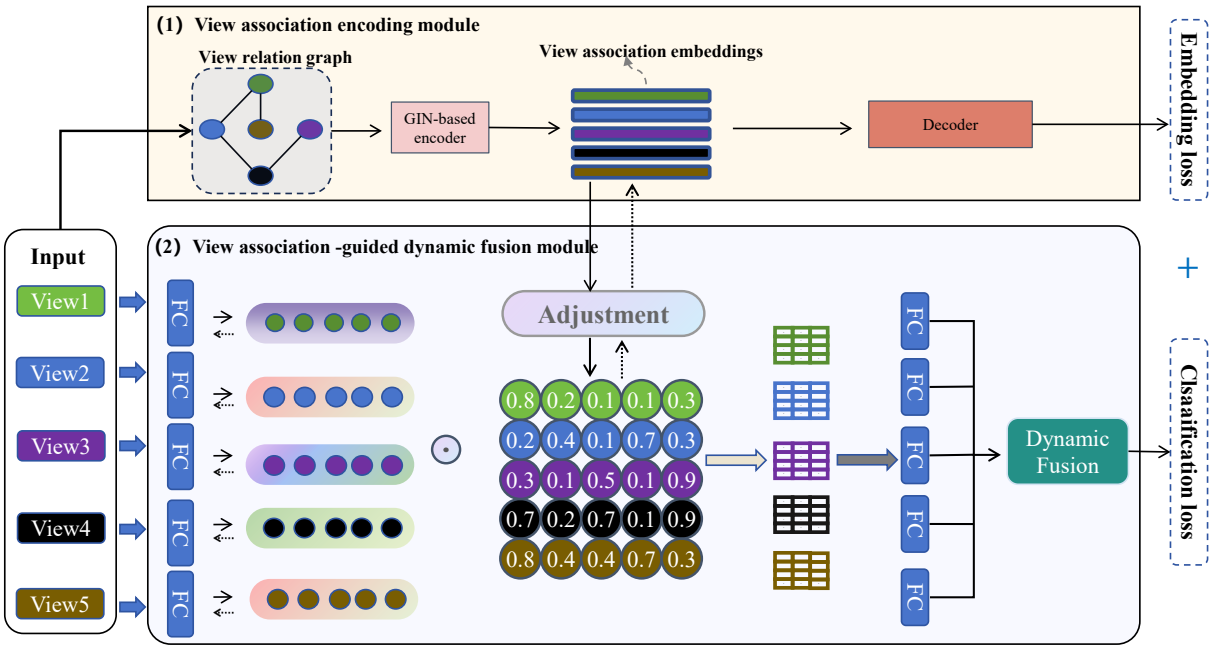

Figure 2: The whole frameword of the view-association-guided dynamic fusion network

(or feature space) of representations from $|V|$ views, where m_i $(1 \leq i \leq |V|)$ denotes the feature dimension of the i-th view, and let $\mathcal{Y} = \{l_1, l_2, \ldots, l_q\}$ represent the label space with q class labels. Let $\mathcal{D}$ denote an unknown distribution over $\mathcal{X} \times \mathcal{Y}$. A training set $\boldsymbol{D} = \{(\boldsymbol{x}_i^v, y_i), |, 1 \leq v \leq |V|, 1 \leq i \leq n\} \in (\mathcal{X} \times \mathcal{Y})^n$ is drawn independently and identically from $\mathcal{D}$, where $\boldsymbol{x}_i^v = (x_{i1}^v, x_{i2}^v, \ldots, x_{im_v}^v) \in \mathbb{R}^{m_v}$ is the feature vector of the v-th view with dimension m_v, and $y_i \in \mathcal{Y}$ is the known label associated with $\boldsymbol{x}_i^v$. The task of multi-view classification is to learn a prediction function $f : \mathcal{X} \to \mathcal{Y}$ from $\boldsymbol{D}$, which can assign an appropriate label $f(\boldsymbol{x}) \in \mathcal{Y}$ to an unseen instance $\boldsymbol{x}$.

3.2 View Association Encoding Module

The module aims to explicitly model the association between different views and learn the corresponding view association embeddings. By capturing the complex between views, the learned view association embeddings enable the model to better understand how each view influences the others, thereby optimizing the fusion process and facilitating more accurate predictions.

First, we construct the view association graph based on the similarity relationships between views. Let $G = (V, E)$ represent the view association graph, where V denotes the set of views and E denotes the set of edges. The adjacency matrix $\mathbf{A}$ stores the weights associated with each edge, representing the similarity between views. Since the feature dimensions of samples in each view are not consistent, the similarity relationships between views cannot be directly calculated. Therefore, modeling the relationships between views is converted to examining the relationships between the features of the same sample across different views. To reduce computational cost, the following strategy is adopted for calculating the similarity between views:

- First, from the same view, select k samples. Let the sample set $\mathcal{X}^v = \{x_1^v, x_2^v, \ldots, x_n^v\}$ be the set of all samples in the v-th view, where each $x_i^v \in \mathbb{R}^{m_v}$ is the feature vector of the i-th sample. Select k samples from these n samples to form a sample set $S_k^v = \{x_{i_1}^v, x_{i_2}^v, \ldots, x_{i_k}^v\}$, where $i_1, i_2, \ldots, i_k \in \{1, 2, \ldots, n\}$.
- Next, calculate the Euclidean distance between each of these k samples and every other sample, resulting in an $n \times k$ Euclidean distance matrix D^v, where the element D_{ij}^v represents the distance between the i-th and j-th samples:

$$D_{ij}^v = \|x_i^v - x_j^v\|_2 = \sqrt{\sum_{m=1}^{m_v} (x_{im}^v - x_{jm}^v)^2}. \quad (1)$$

This matrix D^v is of size $n \times k$, representing the Euclidean distances between each sample and the k selected samples. The obtained Euclidean distance matrix D is subtracted pairwise to get the relative distance matrix R between views:

$$R_{ij} = \frac{1}{n \times k} |D^i - D^j|, \quad (2)$$

where $i, j \in \{1, 2, \ldots, |V|\}$. A smaller relative distance indicates a higher similarity between the views. Finally, we perform a normalization operation on R to obtain the adjacency matrix $\mathbf{A}$.

Subsequently, a graph autoencoder is applied to learn the view association embeddings. The encoder in the graph autoencoder is instantiated by a Graph Isomorphism Network (GIN) . Given a feature matrix $\mathbf{H}^{(t)} \in \mathbb{R}^{|V| \times d^{(t)}}$ for a node,

where each row represents a view association embedding and $d^{(t)}$ denotes the dimension of the node features, the GIN layer updates the node as follows:

$$\mathbf{H}^{(t+1)} = f^{(t+1)}[(1+\epsilon^{(t+1)})\mathbf{H}^{(t)} + \mathbf{A}\mathbf{H}^{(t)}], \tag{3}$$

where $\mathbf{H}^{(t+1)} \in \mathbb{R}^{|V|\times d^{(t+1)}}$ is the updated feature matrix of the nodes, and $f^{(t+1)}$ represents a neural network consisting of two fully connected layers, followed by Batch Normalization and a LeakyReLU activation function. $\epsilon^{(t+1)}$ is a learnable parameter that controls the importance of a node's own features in neighborhood aggregation.

After passing through the GIN layer, $\mathbf{H}^{(T)} \in \mathbb{R}^{|V|\times d^{(T)}}$ is taken as the final view association embedding $\mathbf{E} \in \mathbb{R}^{|V|\times d_e}$, i.e., $\mathbf{E} = \mathbf{H}^{(T)}$. The embedding loss function is:

$$\mathcal{L}_{le} = \frac{1}{V^2}\sum_{i=1}^{|V|}\sum_{j=1}^{|V|}[\cos(e_i, e_j) - \hat{A}_{ij}]^2, \tag{4}$$

where $\cos(e_i, e_j)$ represents the cosine similarity between the embeddings of views V_i and V_j, and $\hat{A} = A + I$, where I is the identity matrix.

3.3 View Association-Guided Dynamic Weighting Fusion Module

View Association-Guided Representation Learning. In multi-view data, each sample in the same dataset has multiple related feature representations which leads to differences in the dimensionality of the feature vectors of samples across different views. This requires unifying the dimensionality of feature vectors from different views to facilitate subsequent computations. Therefore, the first step is to map instances from different views in the original feature space to the same latent space:

$$\boldsymbol{x}_z^v = W\boldsymbol{x}^v + b, \tag{5}$$

where $W \in \mathbb{R}^{d\times m_v}$ and $b \in \mathbb{R}^d$ are learnable parameters.

Then, to enable view association embeddings to better adapt to downstream tasks, an attention-like mechanism, called the Adjustment Module(AM), is designed to guide the sample in focusing on distinct features across different views. Specifically, a two-layer fully connected network is used to transform the view association embeddings into guiding vectors:

$$\alpha_v = \sigma(W_{e2}(W_{e1}e_v + b_{e1}) + b_{e2}), \tag{6}$$

where $\alpha_v \in \mathbb{R}^d$ is the guiding vectors of the v-th view, and σ is the sigmoid function. Successively, the guiding vector is element-wise multiplied (Hadamard product) with the vector after the sample is unified to the same dimension. Subsequently, the selected features are input into their respective learners.

$$z_v = f_v(\boldsymbol{x}_z^v \odot \alpha_v), \tag{7}$$

where f_v represents the learner of the v-th view consisting of one fully connected layer, followed by Batch Normalization and a ReLU activation function.

Dynamic Weighting Fusion. After the feature disentanglement process, the features from each view are adjusted, and their quality may vary depending on the specific characteristics of each view. As a result, the importance and relevance of these features differ across views. Therefore, we introduce Dynamic Uncertainty(DU) to measure the uncertainty of each view [Cao *et al.*, 2024]. The dynamic uncertainty for i-th view can be formulated as follows:

$$DU_i = \sum_{i=1}^{C} \left|\text{Softmax}(z_i) - \mu\right|, \tag{8}$$

where C is the class number, μ is the mean of probability, and it holds $\mu = \frac{1}{C}$. The distribution of probabilities after softmax offers critical insights into a view's uncertainty: A uniform distribution typically suggests high uncertainty, whereas a peaked distribution implies low uncertainty in predictions. In the multi-view fusion process, one view should dynamically perceive the changes of other views and modify its relative contribution to the multi-view system. Therefore, we introduce relative calibration as the weight for multi-view fusion:

$$w_i = \begin{cases} RC_i = \frac{DU_i\cdot(|V|-1)}{\sum_{i\neq m} DU_m} & \text{if } RC_i < 1 \\ 1 & \text{otherwise} \end{cases} \tag{9}$$

Next, the fused features are obtained as follows

$$\boldsymbol{o} = \sum_{i=1}^{|V|} w_i \cdot \boldsymbol{z}_i. \tag{10}$$

Finally, the fused features are passed to a softmax function for computing the class probability as follows:

$$\hat{\boldsymbol{y}} = \text{Softmax}(\boldsymbol{o}). \tag{11}$$

3.4 Overall Objective Function

The AssoDMVC is trained with the following objective function in an end-to-end fashion:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda\mathcal{L}_{\text{le}}, \tag{12}$$

where $\mathcal{L}_{\text{le}}$ is the view association embedding loss, λ is a trade-off parameter, and $\mathcal{L}_{\text{ce}}$ denotes the cross entropy loss, formulated as

$$\mathcal{L}_{\text{ce}} = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij}). \tag{13}$$

4 Experiments

4.1 Datasets

Our experiments are conducted on six challenging multi-view classification datasets which include image, text, audio, depth and video datasets. (1) Animals with Attributes (AWA)[Lampert *et al.*, 2013] dataset, which includes 30,475 images from 50 categories with seven view features. (2) NUS-WIDE-128 (NUS)[Tang *et al.*, 2016] dataset, which includes 43,800 samples from 128 categories with seven view features. (3) Reuters [Amini *et al.*, 2009] dataset, which includes 111,740

Accuracy

Groups	Methods	AWA	NUS	Reuters5	Reuters3	VoxCeleb	YoutubeFace
Feature	EmbraceNet(IF19)	84.97 ± 0.23	72.43 ± 0.38	80.07 ± 0.21	83.58 ± 0.25	81.74 ± 0.34	80.90 ± 1.04
	AWDR(PR19)	$\underline{90.46 \pm 0.06}$	72.44 ± 0.66	79.69 ± 0.27	83.32 ± 0.32	91.08 ± 0.09	85.11 ± 0.15
	RAMC(INS22)	90.63 ± 0.13	72.51 ± 0.67	79.84 ± 0.25	83.48 ± 0.25	91.54 ± 0.11	$\underline{85.21 \pm 0.17}$
Decision	BV(TEVC21)	88.65 ± 0.43	68.69 ± 0.59	80.61 ± 0.25	83.98 ± 0.14	63.25 ± 0.14	82.01 ± 0.18
	SSV(TEVC21)	82.37 ± 1.26	63.70 ± 0.64	79.51 ± 0.41	84.71 ± 0.22	85.10 ± 0.23	84.43 ± 0.31
	MR(TEVC21)	87.10 ± 0.64	64.39 ± 0.85	78.24 ± 0.45	84.17 ± 0.19	79.92 ± 0.29	84.78 ± 0.21
	TMOA(AAAI22)	89.17 ± 0.31	72.60 ± 0.48	79.11 ± 0.43	84.19 ± 0.27	84.72 ± 0.21	84.35 ± 0.25
	TMC(ICLR22)	88.59 ± 0.25	72.73 ± 0.30	79.60 ± 0.56	84.23 ± 0.35	73.13 ± 0.15	71.18 ± 2.27
	ETMC(TPAMI23)	88.24 ± 0.17	73.05 ± 0.67	79.80 ± 0.41	84.24 ± 0.42	88.70 ± 0.15	79.63 ± 1.89
	ECML(AAAI24)	80.51 ± 0.41	72.53 ± 0.55	81.39 ± 0.18	85.88 ± 0.29	89.06 ± 0.21	81.95 ± 0.20
	RMVC(IF25)	81.49 ± 0.31	60.61 ± 0.54	78.89 ± 0.20	82.97 ± 0.19	88.34 ± 0.09	81.56 ± 0.28
	TUNED(AAAI25)	89.05 ± 0.45	$\underline{74.08 \pm 0.36}$	$\underline{81.65 \pm 0.32}$	$\underline{86.02 \pm 0.69}$	$\underline{91.67 \pm 0.30}$	84.79 ± 0.33
Ours	AssoDMVC	$\mathbf{90.86 \pm 0.19}$	$\mathbf{74.62 \pm 0.15}$	$\mathbf{81.79 \pm 0.20}$	$\mathbf{86.04 \pm 0.57}$	$\mathbf{93.85 \pm 0.05}$	$\mathbf{86.21 \pm 0.15}$

Precision

Groups	Methods	AWA	NUS	Reuters5	Reuters3	VoxCeleb	YoutubeFace
Feature	EmbraceNet(IF19)	82.14 ± 0.57	71.73 ± 0.32	80.42 ± 0.25	83.77 ± 0.34	80.95 ± 0.46	83.71 ± 1.10
	AWDR(PR19)	89.32 ± 0.33	72.71 ± 0.61	79.87 ± 0.30	83.49 ± 0.34	91.83 ± 0.11	89.94 ± 0.32
	RAMC(INS22)	$\underline{89.41 \pm 0.38}$	72.82 ± 0.64	80.12 ± 0.27	83.70 ± 0.28	$\underline{92.19 \pm 0.06}$	$\underline{90.64 \pm 0.08}$
Decision	BV(TEVC21)	86.57 ± 0.46	70.98 ± 0.95	80.77 ± 0.19	84.13 ± 0.19	64.63 ± 0.63	84.34 ± 0.61
	SSV(TEVC21)	82.76 ± 1.10	67.23 ± 0.58	80.19 ± 0.49	85.16 ± 0.20	84.44 ± 0.18	$\mathbf{94.13 \pm 0.37}$
	MR(TEVC21)	85.44 ± 0.64	64.90 ± 0.81	78.21 ± 0.48	84.25 ± 0.13	78.85 ± 0.29	86.56 ± 0.58
	TMOA(AAAA22)	88.15 ± 0.62	72.73 ± 0.53	79.89 ± 0.72	84.40 ± 0.23	84.38 ± 0.30	87.59 ± 0.28
	TMC(ICLR22)	87.76 ± 0.40	72.71 ± 0.22	79.86 ± 0.46	84.43 ± 0.49	73.26 ± 0.34	82.53 ± 2.01
	ETMC(TPAMI23)	87.68 ± 0.63	72.39 ± 0.64	79.99 ± 0.33	84.38 ± 0.37	87.28 ± 0.15	83.40 ± 2.33
	ECML(AAAI24)	86.27 ± 1.22	73.05 ± 0.26	$\underline{81.52 \pm 0.18}$	$\underline{85.81 \pm 0.27}$	74.21 ± 0.46	84.34 ± 0.38
	RMVC(IF25)	81.35 ± 0.23	60.09 ± 0.41	80.32 ± 0.45	82.28 ± 0.49	89.26 ± 0.19	81.32 ± 0.23
	TUNED(AAAI25)	89.02 ± 0.12	$\underline{74.12 \pm 0.33}$	81.12 ± 0.22	85.79 ± 0.12	92.01 ± 0.23	85.35 ± 0.15
Ours	AssoDMVC	$\mathbf{89.79 \pm 0.23}$	$\mathbf{75.00 \pm 0.13}$	$\mathbf{82.03 \pm 0.20}$	$\mathbf{85.89 \pm 0.23}$	$\mathbf{93.95 \pm 0.13}$	86.52 ± 0.40

F1

Groups	Methods	AWA	NUS	Reuters5	Reuters3	VoxCeleb	YoutubeFace
Feature	EmbraceNet(IF19)	80.04 ± 0.59	72.04 ± 0.34	79.85 ± 0.26	83.46 ± 0.21	78.36 ± 0.34	80.65 ± 1.13
	AWDR(PR19)	86.86 ± 0.20	71.87 ± 0.62	79.59 ± 0.23	83.30 ± 0.29	87.26 ± 0.13	83.57 ± 0.30
	RAMC(INS22)	87.08 ± 0.42	71.92 ± 0.65	79.73 ± 0.23	83.45 ± 0.23	87.95 ± 0.11	83.35 ± 0.27
Decision	BV(TEVC21)	85.72 ± 0.57	67.67 ± 0.57	80.52 ± 0.29	83.91 ± 0.11	57.79 ± 0.14	81.05 ± 0.35
	SSV(TEVC21)	77.28 ± 1.45	60.52 ± 0.63	79.08 ± 0.40	84.48 ± 0.25	81.07 ± 0.26	80.80 ± 0.53
	MR(TEVC21)	83.55 ± 0.77	63.10 ± 0.91	78.11 ± 0.45	84.12 ± 0.26	75.36 ± 0.32	83.87 ± 0.31
	TMOA(AAAI22)	83.62 ± 0.91	71.81 ± 0.49	78.85 ± 0.30	84.25 ± 0.30	81.54 ± 0.26	82.63 ± 0.39
	TMC(ICLR22)	84.47 ± 0.54	71.70 ± 0.43	79.60 ± 0.56	84.19 ± 0.29	64.06 ± 0.12	68.50 ± 2.77
	ETMC(TPAMI23)	84.60 ± 0.49	72.19 ± 0.68	79.72 ± 0.40	84.24 ± 0.42	86.03 ± 0.20	80.97 ± 1.48
	ECML(AAAI24)	84.82 ± 1.05	72.01 ± 0.52	81.35 ± 0.16	$\underline{85.89 \pm 0.28}$	75.87 ± 0.35	82.30 ± 0.15
	RMVC(IF25)	80.62 ± 0.59	59.97 ± 1.46	79.21 ± 0.85	83.24 ± 0.70	88.14 ± 0.09	81.66 ± 0.15
	TUNED(AAAI25)	$\underline{88.11 \pm 0.73}$	$\underline{73.65 \pm 0.55}$	$\underline{81.46 \pm 0.45}$	85.24 ± 0.44	$\underline{91.84 \pm 0.42}$	$\underline{85.06 \pm 0.57}$
Ours	AssoDMVC	$\mathbf{88.38 \pm 0.34}$	$\mathbf{74.50 \pm 0.15}$	$\mathbf{81.77 \pm 0.15}$	$\mathbf{85.93 \pm 0.20}$	$\mathbf{92.73 \pm 0.11}$	$\mathbf{86.72 \pm 0.40}$

Table 1: Comparison results with SOTA methods. The best and the second best results are highlighted by boldface and underlined respectively.

samples from six categories with five multilingual view features. To enable the model to process this dataset, PCA is used to reduce the dimensions of all views to 1000. According to [Han *et al.*, 2022; Liu *et al.*, 2021], Gaussian noise is added to either 5-view or 3-view datasets, resulting in two versions named Reuters5 and Reuters3, respectively.(4) VoxCeleb [Nagrani *et al.*, 2020] dataset, which includes 153,516 samples from 1,251 categories with five audio view features. (5) YoutubeFace dataset, which includes 3,425 videos from 1,595 different people with five view features. According to [Wang *et al.*, 2022b], we use a subset of 31 categories from this dataset, with a total of 101,499 frames.

We employ three measures to evaluate the performance of each method, which are accuracy, precision and F1 score [Liang *et al.*, 2024].

4.2 Experimental Results with Other Methods

To validate the effectiveness of the our method, comprehensive comparison experiments are conduced with eight related weighting-based multi-view classification methods. The compared methods can be classified into the following two groups according to the level of weighting:

1. The first category is the $feature$ level including EmbraceNet, AWDR and RAMC [Jiang *et al.*, 2022]. EmbraceNet assigns 1 to the weight value of one view while 0 to others for each example according to a multinomial distribution. AWDR is an adaptive-weighting discriminative regression approach. Following [Yang *et al.*, 2019], the parameter λ is chosen from the set $\{10^{-3}, 10^{-2}, \cdots, 10^{3}\}$, while k varies within the range $\{1, 3, \cdots, 9\}$. RAMC employs an $L_{2,1}$-norm loss function to acquire a joint weighted projection space across all views. This method preserves the correlation and diversity among views through a self-supervised weighting strategy. Similarly, the parameter λ is chosen from the set $\{10^{-3}, 10^{-2}, \cdots, 10^{3}\}$ and k ranges from the range $\{1, 3, \cdots, 9\}$.
2. The second category is the $decision$ level including BV, SSV, MR [Liang *et al.*, 2021], TMC [Han *et al.*, 2022], TMOA [Liu *et al.*, 2022], ETMC [Han *et al.*, 2023], ECML [Xu *et al.*, 2024], RMVC [Yue *et al.*, 2025] and TUNED [Huang *et al.*, 2025]. BV assigns 1 to the weight value of the view with the best performance while 0 to others according to whole classification performance of each view. MR assigns 1 to the weight value of the view with the best performance while 0 to others for each example according to the classification performance of each view of each example. SSV assigns the same values to all views. TMC, TMOA, ETMC, ECML and RMVC are trusted fusion methods.

For the proposed AssoDMVC, in order to make the model elegant and lightweight, we set each module to contain one or two fully connected layer, and the number of neurons in the fully connected layer is selected from [128, 256]. We initialize the learnable parameters ϵ in GIN by 0. The results in Tables 1, are presented through the mean metric value and the standard deviation obtained from 5-fold cross-validation. From Table 1, the following observations can be made:

1. The AssoDMVC method shows remarkable performance in accuracy, especially on the VoxCeleb and YoutubeFace datasets, achieving 93.85% and 86.21%, respectively. Additionally, AssoDMVC consistently maintained top or near-top performance across most datasets, including AWA, NUS, Reuters5, and Reuters3, achieving 90.86%, 74.62%, 81.79%, and 86.04%, respectively. The F1 score, which balances precision and recall, provides a comprehensive measure of classification performance. The AssoDMVC shows significant advantages in F1 scores as well, especially on the AWA, NUS, and Reuters5 datasets, where it achieves 88.38%, 74.50%, and 81.77%, respectively, reflecting a good balance between precision and recall.
2. The AssoDMVC shows stable performance across multiple datasets, not only outperforming SOTA methods on some datasets but also matching or coming close to the best performing methods on others. This demonstrates its broad applicability in multi-view learning tasks.

Overall, the AssoDMVC method outperforms existing SOTA methods across several key performance metrics. By introducing new strategies for view relation and dynamic fusion, AssoDMVC handles the complexity and diversity of multi-view data more effectively, demonstrating stronger adaptability and higher model performance.

4.3 Further Analysis

In the part, we further analysis our model from three aspects: each module effectiveness, hyper-parameter sensibility, impact of different weight strategy.

Ablation Experiments. To validate the effectiveness of the components in AssoDMVC, we performed experiments to assess the impact of the view association encoding module (RM), adjustment module (AM), and dynamic weighting (DW) on the experimental results. Results of the ablation study are presented in Table 2.

As shown in the third and fourth rows of the table, adding the RM results in a significant improvement in performance across all datasets. Specifically, the model performs better with the RM, achieving 90.49% on the AWA dataset, which is significantly higher than the 82.27% without it. The AM also has a significant impact on the experimental results. Specifically, the accuracy on the AWA dataset increases from 90.49% to 90.76% when the AM is added. As seen in the first and second rows of the table, adding the DW leads to significant gains. For example, the accuracy on NUS increases from 69.84% to 73.17%. The last row of the table presents the results when all three modules are combined. It is evident that the combination of all three modules yields the superior performance across all datasets.

Parameter Sensitive Analysis. In the View Association Encoding Module, to reduce computational cost, we select k samples for graph construction. To assess the impact of different k values on the experimental accuracy, we perform a comparison using various values of k. The results are shown in the Fig. 3. The accuracy remains relatively consistent across different k values, suggesting that we can select an ap-

RM	AM	DW	AWA	NUS	Reuters5	Reuters3	VoxCeleb	YoutubeFace
×	×	×	82.27	69.84	77.48	82.23	90.06	84.00
×	×	✓	81.35	73.17	78.32	83.74	91.50	85.86
✓	×	×	90.49	74.17	81.12	85.38	93.54	85.93
✓	✓	×	90.76	74.34	81.21	85.20	93.62	86.01
✓	×	✓	90.68	74.09	81.29	85.50	93.55	85.88
✓	✓	✓	**90.86**	**74.62**	**81.79**	**86.04**	**93.85**	**86.21**

Table 2: Ablation results for different components.

propriate k without being overly concerned about its impact on model performance.

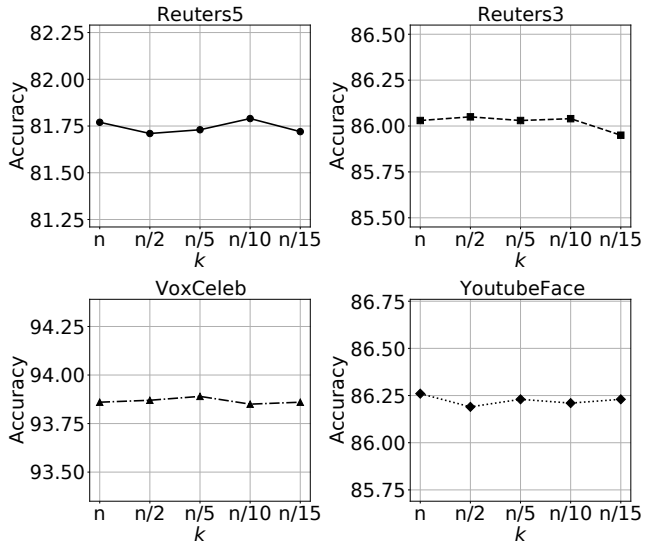

Figure 3: Impact of k values on performance

Different Weight Strategies. We introduced three different weighting strategies for the multi-view fusion process: Average Weighting (Avg), Learnable Weighting (LW), and No Weighting (NW), and compared them with Dynamic Weighting (DW).

1. **Average Weighting (Avg)**: In this strategy, equal weights are assigned to all views, meaning no differentiation is made between the views.
2. **Learnable Weighting (LW)**: This strategy allows the model to learn the optimal weights for each view during the training process.
3. **No Weighting (NW)**: In the No Weighting strategy, no explicit weights are applied to the views.

In the Fig. 4, Dynamic Weighting (DW) demonstrates a clear advantage. Specifically, on the Reuters3 dataset, DW achieved the highest performance at 86.04%, outperforming other weighting strategies such as Average Weighting (Avg), Learnable Weighting (LW), and No Weighting (NW), highlighting its effectiveness and superiority on this dataset. Similarly, on the VoxCeleb dataset, DW also performed well, reaching 93.85%, showing a slight improvement over other methods.

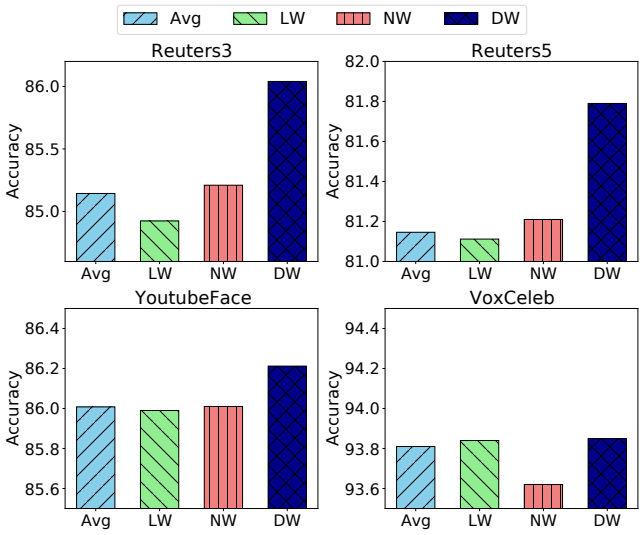

Figure 4: Impact of different weight strategies on performance

Visualization. To provide deeper insights into AssoDMVC, we visualize the learned view association embeddings on the AWA dataset. As shown in Fig. 5, views 0, 2, 3, and 4 exhibit notably high correlations, suggesting strong mutual dependencies among these views. These results indicate that AssoDMVC effectively guides multi-view fusion by modeling the associations between views.

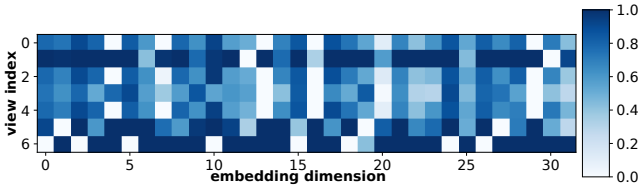

Figure 5: Visualization of AssoDMVC on AWA

5 Conclusion

In this paper, we proposed a view-association-guided dynamic multi-view classification method, which effectively models the relationships between views and dynamically adjusts their contributions. By introducing a view association encoding module and a relative calibration mechanism, our method enhances view interaction and improves fusion accuracy. Experimental results show that our approach outperforms traditional fusion techniques, providing a more robust and flexible solution for multi-view learning tasks. Future work could explore more effective view interaction mechanisms to further enhance the information fusion and collaboration between views.

Acknowledgements

This work was supported by National Natural Science Foundation of China (Nos. 62306171, 62406218, T2495251,

T2495253, 62476160, 62376146, 62373233), the Science and Technology Major Project of Shanxi (No. 202201020101006), and Fundamental Research Program of Shanxi Province (No. 202203021222183).

References

[Amini *et al.*, 2009] Massih R Amini, Nicolas Usunier, and Cyril Goutte. Learning from multiple partially observed views-an application to multilingual text categorization. *Advances in Neural Information Processing Systems*, 22, 2009.

[Cao *et al.*, 2024] Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Predictive dynamic fusion. In *Forty-first International Conference on Machine Learning*, 2024.

[Chen *et al.*, 2021] Yongyong Chen, Xiaolin Xiao, Zhongyun Hua, and Yicong Zhou. Adaptive transition probability matrix learning for multiview spectral clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4712–4726, 2021.

[Chen *et al.*, 2023] Zhaoliang Chen, Lele Fu, Jie Yao, Wenzhong Guo, Claudia Plant, and Shiping Wang. Learnable graph convolutional network and feature fusion for multiview learning. *Information Fusion*, 95:109–119, 2023.

[Choi and Lee, 2019] Jun-Ho Choi and Jong-Seok Lee. Embracenet: A robust deep learning architecture for multimodal classification. *Information Fusion*, 51:259–270, 2019.

[Fu *et al.*, 2024a] Pinhan Fu, Xinyan Liang, Tingjin Luo, Qian Guo, Yayu Zhang, and Yuhua Qian. Core-structures-guided multi-modal classification neural architecture search. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 3980–3988, 2024.

[Fu *et al.*, 2024b] Pinhan Fu, Xinyan Liang, Yuhua Qian, Qian Guo, Zhifang Wei, and Wen Li. Como-nas: Core-structures-guided multi-objective neural architecture search for multi-modal classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 9126–9135, 2024.

[Fu *et al.*, 2025] Pinhan Fu, Xinyan Liang, Yuhua Qian, Qian Guo, Yayu Zhang, Qin Huang, and Ke Tang. Multi-scale features are effective for multi-modal classification: An architecture search viewpoint. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2):1070–1083, 2025.

[Guo *et al.*, 2024] Qian Guo, Xinyan Liang, Yuhua Qian, Zhihua Cui, and Jie Wen. A progressive skip reasoning fusion method for multi-modal classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 429–437, 2024.

[Han *et al.*, 2022] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, pages 1–11, 2022.

[Han *et al.*, 2023] 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, 2023.

[Huang *et al.*, 2020] Aiping Huang, Tiesong Zhao, and Chia-Wen Lin. Multi-view data fusion oriented clustering via nuclear norm minimization. *IEEE Transactions on Image Processing*, 29:9600–9613, 2020.

[Huang *et al.*, 2025] 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*, 2025.

[Jiang *et al.*, 2021] Bingbing Jiang, Junhao Xiang, Xingyu Wu, Wenda He, Libin Hong, and Weiguo Sheng. Robust adaptive-weighting multi-view classification. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, page 3117–3121, 2021.

[Jiang *et al.*, 2022] Bingbing Jiang, Junhao Xiang, Xingyu Wu, Yadi Wang, Huanhuan Chen, Weiwei Cao, and Weiguo Sheng. Robust multi-view learning via adaptive regression. *Information Sciences*, 610:916–937, 2022.

[Jiang *et al.*, 2024] Zhangqi Jiang, Tingjin Luo, and Xinyan Liang. Deep incomplete multi-view learning network with insufficient label information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11):12919–12927, 2024.

[Lampert *et al.*, 2013] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2013.

[Li and Tang, 2024] Songtao Li and Hao Tang. Multimodal alignment and fusion: A survey. *arXiv preprint arXiv:2411.17040*, 2024.

[Li *et al.*, 2023] Guangyao Li, Wenxuan Hou, and Di Hu. Progressive spatio-temporal perception for audio-visual question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7808–7816, 2023.

[Liang *et al.*, 2021] Xinyan Liang, Qian Guo, Yuhua Qian, Weiping Ding, and Qingfu Zhang. Evolutionary deep fusion method and its application in chemical structure recognition. *IEEE Transactions on Evolutionary Computation*, 25(5):883–893, 2021.

[Liang *et al.*, 2022] Xinyan Liang, Yuhua Qian, Qian Guo, Honghong Cheng, and Jiye Liang. AF: An association-based fusion method for multi-modal classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9236–9254, 2022.

[Liang *et al.*, 2024] Xinyan Liang, Pinhan Fu, Qian Guo, Keyin Zheng, and Yuhua Qian. DC-NAS: Divide-and-conquer neural architecture search for multi-modal clas-

sification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13754–13762, 2024.

[Liang *et al.*, 2025] Xinyan Liang, Pinhan Fu, Yuhua Qian, Qian Guo, and Guoqing Liu. Trusted multi-view classification via evolutionary multi-view fusion. In *Proceedings of the 13th International Conference on Learning Representations*, pages 1–14, 2025.

[Liu *et al.*, 2021] Bo Liu, Haowen Zhong, and Yanshan Xiao. New multi-view classification method with uncertain data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(1):1–23, 2021.

[Liu *et al.*, 2022] Wei Liu, Xiaodong Yue, Yufei Chen, and Thierry Denoeux. Trusted multi-view deep learning with opinion aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7585–7593, 2022.

[Nagrani *et al.*, 2020] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027, 2020.

[Tang *et al.*, 2016] Jinhui Tang, Xiangbo Shu, Guo-Jun Qi, Zechao Li, Meng Wang, Shuicheng Yan, and Ramesh Jain. Tri-clustered tensor completion for social-aware image tag refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1662–1674, 2016.

[Wang *et al.*, 2022a] Shiping Wang, Zhaoliang Chen, Shide Du, and Zhouchen Li. Learning deep sparse regularizers with applications to multi-view clustering and semi-supervised classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5042–5055, 2022.

[Wang *et al.*, 2022b] Siwei Wang, Xinwang Liu, Li Liu, Wenxuan Tu, Xinzhong Zhu, Jiyuan Liu, Sihang Zhou, and En Zhu. Highly-efficient incomplete large-scale multi-view clustering with consensus bipartite graph. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9776–9785, 2022.

[Wen *et al.*, 2023] Jie Wen, Zheng Zhang, Lunke Fei, Bob Zhang, Yong Xu, Zhao Zhang, and Jinxing Li. A survey on incomplete multiview clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(2):1136–1149, 2023.

[Xu *et al.*, 2024] 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*, volume 38, pages 16129–16137, 2024.

[Yang *et al.*, 2019] Muli Yang, Cheng Deng, and Feiping Nie. Adaptive-weighting discriminative regression for multi-view classification. *Pattern Recognition*, 88:236–245, 2019.

[Yue *et al.*, 2025] Xiaodong Yue, Zhicheng Dong, Yufei Chen, and Shaorong Xie. Evidential dissonance measure in robust multi-view classification to resist adversarial attack. *Information Fusion*, 113:102605, 2025.

[Zhang *et al.*, 2024a] Chenglong Zhang, Yang Fang, Xinyan Liang, Han Zhang, Peng Zhou, Xingyu Wu, Jie Yang, Bingbing Jiang, and Weiguo Sheng. Efficient multi-view unsupervised feature selection with adaptive structure learning and inference. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*, pages 5443–5452, 2024.

[Zhang *et al.*, 2024b] Chenglong Zhang, Xinjie Zhu, Zidong Wang, Yan Zhong, Weiguo Sheng, Weiping Ding, and Bingbing Jiang. Discriminative multi-view fusion via adaptive regression. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.