

# Trusted Unified Feature-Neighborhood Dynamics for Multi-View Classification

Haojian Huang<sup>1, 2</sup>, Chuanyu Qin<sup>3</sup>, Zhe Liu<sup>4</sup>, Kaijing Ma<sup>1, 5</sup>, Jin Chen<sup>1, 5</sup>, Han Fang<sup>1</sup>, Chao Ban<sup>1</sup>,  
Hao Sun<sup>1\*</sup>, Zhongjiang He<sup>1\*</sup>

<sup>1</sup>TeleAI, China

<sup>2</sup>The University of Hong Kong, Hong Kong, China

<sup>3</sup>Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

<sup>4</sup>Universiti Sains Malaysia, Penang, Malaysia

<sup>5</sup>Xi'an Jiaotong University, Xi'an, China

haojianhuang@connect.hku.hk, qinchuan24@mails.ucas.ac.cn,  
sun.010@163.com, hezhongj\_1@163.com

## Abstract

Multi-view classification (MVC) faces inherent challenges due to domain gaps and inconsistencies across different views, often resulting in uncertainties during the fusion process. While Evidential Deep Learning (EDL) has been effective in addressing view uncertainty, existing methods predominantly rely on the Dempster-Shafer combination rule, which is sensitive to conflicting evidence and often neglects the critical role of neighborhood structures within multi-view data. To address these limitations, we propose a Trusted Unified Feature-NEighborhood Dynamics (TUNED) model for robust MVC. This method effectively integrates local and global feature-neighborhood (F-N) structures for robust decision-making. Specifically, we begin by extracting local F-N structures within each view. To further mitigate potential uncertainties and conflicts in multi-view fusion, we employ a selective Markov random field that adaptively manages cross-view neighborhood dependencies. Additionally, we employ a shared parameterized evidence extractor that learns global consensus conditioned on local F-N structures, thereby enhancing the global integration of multi-view features. Experiments on benchmark datasets show that our method improves accuracy and robustness over existing approaches, particularly in scenarios with high uncertainty and conflicting views.

**Code** — <https://github.com/JethroJames/TUNED>

## Introduction

Multi-view classification (MVC) has become a prominent research area due to the increasing availability of data from diverse sources. However, the existing MVC approaches predominantly encode multi-view features into a low-dimensional feature space through a feature bottleneck for fusion and classification. This straightforward fusion process often encounters inherent challenges, such as domain gaps and varying information content across views, which can introduce latent uncertainties during the fusion stage. To address this, many methods have explored measuring the quality of individual view representations to achieve discriminative multi-view fusion. Notably, Evidential Deep

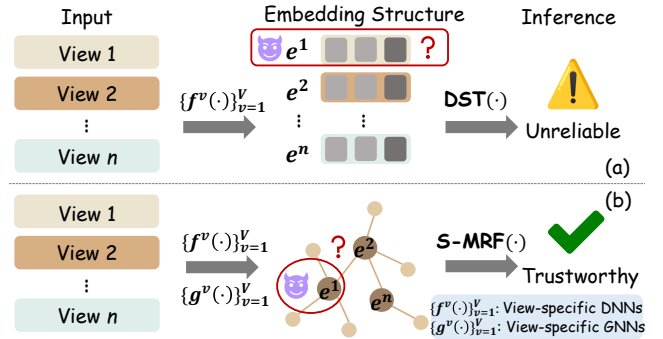


Figure 1: Comparison of EDL-based methods. (a) Conventional models use DNNs  $\{f^v(\cdot)\}_{v=1}^V$  for evidence extraction, neglecting view-specific neighborhood structures and relying on the Dempster-Shafer theory (DST) for direct fusion, which can be unreliable with conflicting views. (b) Our method incorporates feature-neighborhood information via DNNs  $\{f^v(\cdot)\}_{v=1}^V$  and GNNs  $\{g^v(\cdot)\}_{v=1}^V$ . The proposed Selective Markov Random Field (S-MRF) module dynamically fuses evidence, improving inference reliability without the need for hand-crafted loss functions.

Learning (EDL), which stems from subjective logic (Jøsang 2016) and Dempster-Shafer theory (Dempster 1968), has demonstrated effectiveness in assessing view representation uncertainty across various studies (Huang et al. 2024b; Bao, Yu, and Kong 2021; Yu et al. 2023; Qin et al. 2022; Shao, Dou, and Pan 2024; Xu et al. 2024; Yue et al. 2024).

Nevertheless, current EDL-based multi-view learning methods heavily rely on the Dempster-Shafer fusion framework. While theoretically robust, this framework is highly sensitive to conflicting evidence, where even a single contradictory source can produce anomalous results, ultimately compromising inference performance (Huang et al. 2023a; Xiao 2019). Additionally, these methods often overlook the critical importance of neighborhood structure in multi-view features, leading to suboptimal performance in downstream tasks. For instance, in facial expression recognition, focusing solely on isolated view features can lead to misinterpretations due to variations such as lighting or angles. Without effectively leveraging both local and global neighborhood

\*Corresponding author

relationships, models struggle to capture the full context of the data, resulting in a less reliable fusion process. Unfortunately, existing EDL-based methods do not effectively leverage both local and global neighborhood relationships, which hinders their ability to fully capture the context of the data and leads to a less reliable fusion process.

To address these challenges, we propose a model for robust MVC method called Trusted Unified Feature-Neighborhood Dynamics (TUNED). To be specific, in the feature extraction stage, our method not only captures view-specific information but also integrates local neighborhood structures within each view. During the fusion stage, we simultaneously consider global neighborhood structures and the learning of cross-view dependencies, allowing for a more comprehensive understanding of the relationships between different views. To further refine this process, we introduce a dynamic Markov random field that adaptively accounts for cross-view relationships, suppressing views with lower coherence and learning optimal weights for each view. Moreover, We employ a parameterized evidence extractor to learn a collaborative evidence distribution. From this distribution, we sample evidence and perform a locally conditioned fusion with the evidence from each view. This approach enhances the joint learning of local and global feature-neighborhood structures, facilitating more effective and conflict-resilient integration across multiple views. Our contributions are threefold as follows:

- We introduce a novel MVC framework that seamlessly integrates both local and global neighborhood structures. This approach enhances the robustness of feature extraction and fusion processes by effectively capturing cross-view dependencies and mitigating potential conflicts.
- We develop a selective Markov random field (S-MRF) model combined with a parameterized evidence extractor, which adaptively learns and fuses collaborative evidence across multiple views. This dual-layered evidence integration significantly improves the model’s ability to handle complex, heterogeneous data sources.
- Extensive experiments on benchmark datasets demonstrate that our method consistently outperforms state-of-the-art approaches, achieving superior accuracy and robustness in MVC tasks, particularly in scenarios with high levels of uncertainty and conflicts.

## Related Work

### Conflictive multi-view learning

Most conflictive multi-view learning approaches focus on eliminating conflicting data instances. One line of work is based on multi-view outlier detection, which identifies outliers exhibiting abnormal behavior across multiple views. These methods are generally categorized into two types: cluster-based (Huang et al. 2023b; Zhang et al. 2023; Liu et al. 2024) and self-representation-based (Wang et al. 2019; Wen et al. 2023b). Cluster-based methods perform clustering independently in each view and generate affinity vectors for each instance (Marcos Alvarez et al. 2013; Zhao et al.

2017). Outliers are then detected by comparing these vectors across different views. In contrast, self-representation-based methods detect outliers by evaluating their difficulty in being represented by the normal views (Hou et al. 2020).

Another approach focuses on partially view-aligned multi-view learning (Zhang et al. 2021; Wen et al. 2023c). Early work (Lampert and Krömer 2010) introduced weakly-paired maximum covariance analysis to address challenges posed by unaligned data. More recently, (Huang et al. 2020) employed a differentiable variant of the Hungarian algorithm to align unaligned data. Subsequent work proposed noise-robust contrastive learning (Yang et al. 2021; Li 2022; Qin et al. 2024) to compute alignment matrices.

However, these conflictive multi-view learning methods primarily aim to eliminate conflicting instances to achieve coordinated decisions across views. In scenarios like adversarial attack defense, making careful decisions in the presence of conflicting instances is crucial. To address this, recent methods have leveraged Evidential Deep Learning’s (EDL) uncertainty measurement capabilities, quantifying both the conflict degree of among global proxy views (Liu et al. 2023, 2022; Xu et al. 2024) or local intra-view dissonance (Yue et al. 2024). These methods seek to coordinate multiple views for decision-making without discarding conflicting instances. However, they universally rely on the Dempster-Shafer evidence fusion framework, which is highly sensitive to conflicting views (Xiao 2019; Huang et al. 2023a). Without hand-crafted loss functions, this framework often yields counterintuitive and unreliable results, degrading task performance. To overcome these limitations, we propose a novel multi-view fusion framework that introduces a global-local F-N joint learning paradigm, further mitigating multi-view conflicts and enabling more reliable decision-making.

### Uncertainty-based deep learning

Deep neural networks have made significant strides across various tasks, yet they often struggle to quantify uncertainty, especially when dealing with noisy or low-quality data (Wen et al. 2023a; Liu, Huang, and Letchmunan 2023; Liu et al. 2024; Chen et al. 2024a,b). To address these uncertainties, several approaches have been developed, including deterministic methods (Sensoy, Kaplan, and Kandemir 2018), Bayesian neural networks (Gal and Ghahramani 2016), ensemble methods (Lakshminarayanan, Pritzel, and Blundell 2017), and test-time augmentation (Lyzhov et al. 2020). These methods offer different mechanisms for uncertainty estimation, either by enhancing model robustness through probabilistic modeling or by using multiple predictions or augmentations for more accurate uncertainty quantification.

Recently, EDL, a prominent deterministic approach evolving from subjective logic (Jøsang 2016) and Dempster-Shafer theory (DST) (Dempster 1968), has shown promise in uncertainty estimation. EDL has been successfully applied to tasks like MVC (Han et al. 2021, 2022; Liu et al. 2022, 2023; Xu et al. 2024; Huang et al. 2024a), zero-shot learning (Huang et al. 2024b), cross-modal retrieval (Qin et al. 2022), and video understanding (Chen, Gao, and Xu 2023; Gao, Chen, and Xu 2023; Ma et al. 2024; Chen et al.

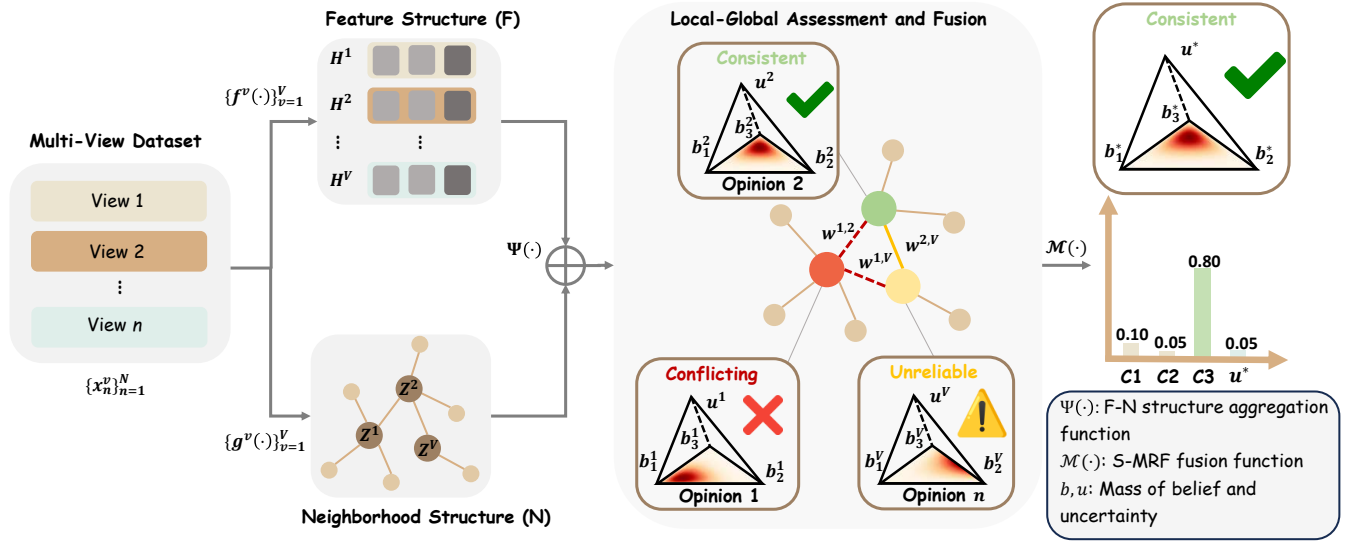


Figure 2: Illustration of the TUNED workflow, which comprises two stages. In the first stage, local view-specific feature-neighborhood (F-N) structures are extracted and fused to obtain evidence from the multi-view dataset. In the second stage, this evidence is integrated with global consensus F-N features through joint learning. The final step involves local-global evidence evaluation via EDL, followed by evidence fusion using the S-MRF module, leading to reliable inference outcomes.

2024c), demonstrating its versatility. In MVC, the pioneering work (Han et al. 2021) introduced the Dempster-Shafer combination rule, sparking a series of EDL-based studies. However, most of these methods assume that uncertainty will decrease after evidence fusion. The Dempster-Shafer rule, being highly sensitive to conflicting views, can produce counterintuitive and unreliable results when faced with conflicting evidence (Xiao 2019; Huang et al. 2023a). Despite recent efforts to improve global coordination and local consistency through hand-crafted loss functions, the limitations of the Dempster-Shafer rule still hinder EDL’s progress in MVC. To this end, we propose a novel fusion framework that mitigates the impact of conflicting views without relying on complex similarity measures or intricate loss designs.

## Methodology

This section presents our proposed method for enhancing EDL through robust evidence fusion in conflictive MVC using the S-MRF. We begin by formally defining the problem, followed by a detailed explanation of how our model incorporates view-specific local F-N structures during the evidence extraction phase. Subsequently, we describe the use of S-MRF to suppress inter-view conflict dependencies, enabling the joint learning of local-global F-N structures. We then outline the loss functions involved in our approach. Finally, we discuss the factors contributing to the effectiveness of our method.

### Problem Definition

In MVC, we are given a dataset  $\mathcal{D} = \{\mathcal{X}_i, \mathbf{y}_i\}_{i=1}^n$ , where  $n$  denotes the number of samples. Each sample  $\mathcal{X}_i$  consists of multiple views, represented as  $\mathcal{X}_i = \{\mathbf{x}_i^v\}_{v=1}^V$ , where  $V$  is the total number of views, and  $\mathbf{x}_i^v \in \mathbb{R}^{d_v}$  represents the

feature vector of the  $i$ -th sample in the  $v$ -th view. The corresponding label for each sample is denoted by  $\mathbf{y}_i \in \mathcal{Y}$ , where  $\mathcal{Y}$  is the set of possible labels. The target of MVC is to learn a model  $F_\theta : \mathcal{X} \rightarrow \mathcal{Y}$  that accurately predicts the label  $\mathbf{y}_i$  for an unseen sample  $\mathcal{X}_i$  by effectively integrating information from all available views  $\{\mathbf{x}_i^v\}_{v=1}^V$ . The challenge lies in leveraging the complementary information from each view while handling potential inconsistencies or conflicts among the views, ultimately improving the overall classification performance.

### Local F-N structure extraction

**Local feature structure embedding.** We simply extract local features from each view using view-specific DNNs. Specifically, for each view  $v$ , we extract feature representations  $\mathbf{h}_n^v$  from the input data using DNNs  $\{f^v(\cdot)\}_{v=1}^V$ . Given an input feature vector  $\mathbf{x}_n^v$ , the corresponding DNN  $f^v(\cdot)$  maps this to a feature vector:

$$\mathbf{h}_n^v = f^v(\mathbf{x}_n^v; \theta^v) \quad (1)$$

where  $\theta^v$  denotes the parameters of the DNN for view  $v$ . These local feature representations inherently capture the local neighborhood structure within each view, providing critical information for subsequent integration.

**Local neighborhood structure embedding.** In multi-view data analysis, traditional feature extraction methods often fail to account for the inherent local neighborhood structures among samples. This omission is particularly problematic in datasets with non-Euclidean characteristics, such as social or biological networks, where intricate patterns are prevalent. Capturing both local and global relationships is crucial for fully understanding the data’s complexity. To effectively capture the complex local neighborhood relation-

ships within multi-view data, we employ the Clustering-with-Adaptive-Neighbors (CAN) (Nie, Wang, and Huang 2014) approach to construct an adaptive adjacency matrix for our Graph Convolutional Network (GCN). This method dynamically determines the neighborhood structure for each sample, enabling the model to robustly reflect both local and global data characteristics.

Given a feature matrix  $X^v \in \mathbb{R}^{d_v \times n}$ , where  $d_v$  represents the feature dimension for view  $v$  and  $n$  represents the number of samples, we first compute the pairwise distance matrix  $D \in \mathbb{R}^{n \times n}$ , where each element  $D_{ij} = \|\mathbf{x}_i^v - \mathbf{x}_j^v\|_2^2$  represents the squared distance between samples  $\mathbf{x}_i^v$  and  $\mathbf{x}_j^v$ . Next, we sort the distances for each sample to identify its top  $k$  nearest neighbors. Let  $\varphi(\mathbf{x}_i^v)$  denote the distance to the  $k$ -th nearest neighbor for sample  $\mathbf{x}_i^v$ . We then compute a weight matrix  $\mathcal{M}$  using the formula  $T_{ij} = \varphi(\mathbf{x}_i^v) - D_{ij}$ , which balances the influence of each neighbor as follows:

$$\mathcal{M}_{ij} = \max\left(0, \frac{T_{ij}}{k \cdot \varphi(\mathbf{x}_i^v) - \sum_{j \in N_i} D_{ij}}\right), \quad (2)$$

where  $N_i$  represents the set of  $k$  nearest neighbors for sample  $\mathbf{x}_i^v$ . This adaptive weighting strategy ensures that each sample's neighborhood captures the most relevant local relationships, leading to a more representative and robust graph structure. The symmetrized adjacency matrix  $\mathbf{A}$  is then computed as:

$$\mathbf{A}^v = \frac{\mathcal{M}^v + (\mathcal{M}^v)^\top}{2}, \quad (3)$$

where  $\mathbf{A}$  is used as the input for the view-specific GCN, and the feature extraction process is defined as below:

$$\mathbf{Q}_{l+1}^v = \sigma(\hat{\mathbf{A}}^v \mathbf{Q}_l^v \mathbf{W}_l^v), \quad (4)$$

where  $\hat{\mathbf{A}}^v$  is the normalized adjacency matrix,  $\sigma$  is the activation function (e.g., ReLU, Softplus), and  $\mathbf{W}_l^v$  are the learnable weights at layer  $l$ .

## Neighborhood-aware evidential deep learning

In multi-view learning, capturing the complex relationships among data points is essential for accurate classification. Traditional methods often overlook neighborhood structures, leading to suboptimal integration of multi-view data, especially in non-Euclidean spaces like social or biological networks. To address this, we propose neighborhood-aware EDL, which explicitly incorporates neighborhood structure. By fusing local-global F-N structures, our approach enhances the model's ability to account for dependencies among data points and achieve more reliable and robust classification.

**Local evidence extraction.** In our framework, the neighborhood-aware features are fused using a function  $\Psi(\cdot)$  (See Supplementary Material) with an activation layer (e.g., ReLU, Softplus) to obtain non-negative output as evidence for each view, so that we can obtain the parameters of the Dirichlet distribution. For a classification problem with  $K$  classes, the evidence  $e_n^v$  for an instance  $\mathbf{x}_n^v$  in view  $v$  is represented by an opinion  $\omega^v = (\mathbf{b}^v, u^v)$ , where  $\mathbf{b}^v =$

$(b_1^v, \dots, b_K^v)^\top$  is the belief mass vector, which assigns belief values to each class based on the evidence and  $u$  is the uncertainty mass, representing the degree of uncertainty in the evidence. These components are constrained by the following relationship:

$$\sum_{k=1}^K b_k^v + u^v = 1, \quad \text{where } b_k^v \geq 0 \quad \text{and} \quad u^v \geq 0. \quad (5)$$

The belief  $\mathbf{b}^v$  and uncertainty  $u^v$  are linked to the Dirichlet distribution, which models second-order uncertainty. The Dirichlet probability density function (PDF) is defined as:

$$D(\mathbf{p}|\boldsymbol{\alpha}^v) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha}^v)} \prod_{k=1}^K p_k^{\alpha_k^v - 1}, & \text{for } \mathbf{p} \in \mathcal{S}_K^v, \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

where  $\mathbf{p}^v = (p_1^v, \dots, p_K^v)^\top$  is the probability assigned to each class  $k$ , and  $\boldsymbol{\alpha}^v = (\alpha_1, \dots, \alpha_K)^\top$  are the Dirichlet parameters. The simplex  $\mathcal{S}_K$  is defined as:

$$\mathcal{S}_K^v = \left\{ \mathbf{p}^v \mid \sum_{k=1}^K p_k^v = 1, \quad 0 \leq p_k^v \leq 1 \text{ for all } k \right\}. \quad (7)$$

The view-specific Dirichlet parameters  $\boldsymbol{\alpha}^v$  are computed as  $\boldsymbol{\alpha}^v = \mathbf{e}^v + \mathbf{1}$ , ensuring non-sparsity. The belief and uncertainty are related to these parameters by:

$$b_k^v = \frac{e_k^v}{S^v} = \frac{\alpha_k^v - 1}{S^v}, \quad u^v = \frac{K}{S^v}, \quad (8)$$

where  $S^v = \sum_{k=1}^K (e_k^v + 1) = \sum_{k=1}^K \alpha_k^v$ . The uncertainty  $u$  is inversely proportional to the total evidence, indicating that as the evidence increases, the uncertainty decreases.

After obtaining the view-specific evidence, this step reflects the confidence of the local proxy in classifying based on the F-N structures. This evidence is then further integrated with global F-N structures in a joint learning manner, setting the stage for robust multi-view evidences fusion.

**Global consensus evidence extraction.** To resolve conflicts between views and establish a robust multi-view consensus, we introduce a consensus evidence extraction mechanism. This mechanism samples a consensus evidence vector  $\mathbf{e}_n^{\text{cons}}$  from a shared Dirichlet distribution and conditions it on the local F-N structures through a fusion function  $\Phi(\cdot)$ , which integrates the sampled consensus evidence with the view-specific evidence vectors as follows:

$$\tilde{\mathbf{e}}_n^v = \mathbf{e}_n^v + \Phi(\mathbf{e}_n^\diamond, \mathbf{e}_n^v), \quad (9)$$

where  $\mathbf{e}_n^\diamond \sim \text{Dir}(\boldsymbol{\alpha}_n^\diamond)$  is the evidence vector randomly sampled from the shared Dirichlet distribution parameterized by  $\boldsymbol{\alpha}_n^\diamond$ . The fusion function  $\Phi(\cdot)$  (e.g. Self-Attention, cosine similarity *etc.*) conditions the consensus evidence on the local F-N structures, facilitating robust global learning of multi-view consensus evidence. This approach ensures that the consensus evidence effectively reflects patterns across all views, while also being adaptable to local variations.

## Conflicting evidence aggregation

Conflicts among views are prevalent yet often overlooked by existing methods. Directly integrating all views can lead to performance degradation, as seen in frameworks like the Dempster-Shafer Theory, where minor local conflicts can cause global fusion failures. To address this, we propose the S-MRF model, which adaptively selects highly relevant views for evidence fusion. This model mitigates the impact of local conflicts on global neighborhood-aware evidence integration, ensuring robust and reliable inference.

**Selective Markov random field.** Specifically, the S-MRF model represents the interaction among views through an energy function, which considers both unary and pairwise potentials. The energy function is defined as below:

$$E(\mathbf{Z}) = \sum_i \phi(\mathbf{z}_i) + \sum_{i,j} \sigma(\mathbf{z}_i, \mathbf{z}_j), \quad (10)$$

where  $\phi(\mathbf{z}_i)$  represents the unary potential representation of view  $i$ , and  $\sigma(\mathbf{z}_i, \mathbf{z}_j)$  represents the pairwise potential representation between views  $i$  and  $j$ . In the S-MRF model, each view  $v$  is associated with an evidence vector  $\mathbf{e}_v$ . A graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  is constructed, where  $\mathcal{V}$  corresponds to the set of views, and  $\mathcal{E}$  represents the set of edges determined by the similarity between views. The edge weight  $w_{i,j}$  reflects the similarity between views  $i$  and  $j$ , calculated using metrics such as cosine similarity or other suitable measures.

Edges are included in the graph if they meet a threshold  $\tau$ , ensuring that only significant and relevant connections are retained, thus addressing inter-view conflicts:

$$w_{i,j} \geq \tau \cdot w_{\max}, \quad (11)$$

where  $w_{\max}$  is the maximum edge weight in the initial connectivity matrix.

**Local-global evidence aggregation.** According to equation (9), we denote view  $v$  evidence as  $\tilde{\mathbf{E}}_v$ , which accounts for local-global F-N structures. The aggregated evidence  $\tilde{\mathbf{E}}_{\text{agg}}$  from these selectively connected views is computed as:

$$\tilde{\mathbf{E}}_{\text{agg}} = \sum_{(i,j) \in \mathcal{E}} w_{i,j} \cdot \tilde{\mathbf{E}}_j. \quad (12)$$

By normalizing the contributions, the S-MRF model effectively combines evidence from views that are contextually significant, reducing the impact of irrelevant or conflicting views on the overall integration process. This selective approach enhances integration efficiency and improves the interpretability of the resulting evidence fusion.

## Loss function

**Neighborhood-aware EDL loss.** For each instance  $\{\mathbf{x}_n^v\}_{v=1}^V$ , we fuse the extracted feature structure  $\mathbf{h}_n^v$  and neighborhood structure  $\mathbf{q}_n^v$  using a F-N aggregation function  $\Psi(\cdot)$  to obtain evidence  $\mathbf{e}_n^v$  for each view:

$$\mathbf{e}_n^v = \Psi(\mathbf{h}_n^v, \mathbf{q}_n^v). \quad (13)$$

This evidence vector  $\mathbf{e}_n^v$  is then used to parameterize the Dirichlet distribution  $\boldsymbol{\alpha}_n^v = \mathbf{e}_n^v + \mathbf{1}$ . Then we adapt the conventional cross-entropy loss to integrate the evidence and

uncertainty from each view, enhancing the robustness of the classification process as follows.

$$\begin{aligned} \mathcal{L}_{acc}(\boldsymbol{\alpha}_n^v) &= \int \left[ \sum_{j=1}^K -y_{nj} \log p_{nj} \right] \frac{\prod_{j=1}^K p_{nj}^{\alpha_{nj}^v - 1}}{B(\boldsymbol{\alpha}_n^v)} d\mathbf{p}_n \\ &= \sum_{j=1}^K y_{nj} (\psi(S_n^v) - \psi(\alpha_{nj}^v)), \end{aligned} \quad (14)$$

where  $\psi(\cdot)$  is the digamma function. The above loss function does not guarantee that the evidence generated by the incorrect labels is lower. To address this issue, we can introduce an additional term in the loss function, namely the Kullback-Leibler (KL) divergence:

$$\begin{aligned} \mathcal{L}_{KL}(\boldsymbol{\alpha}_n) &= KL [D(\mathbf{p}_n | \tilde{\boldsymbol{\alpha}}_n) \| D(\mathbf{p}_n | \mathbf{1})] \\ &= \log \left( \frac{\Gamma(\sum_{k=1}^K \tilde{\alpha}_{nk})}{\Gamma(K) \prod_{k=1}^K \Gamma(\tilde{\alpha}_{nk})} \right) \\ &\quad + \sum_{k=1}^K (\tilde{\alpha}_{nk} - 1) \left[ \psi(\tilde{\alpha}_{nk}) - \psi\left(\sum_{j=1}^K \tilde{\alpha}_{nj}\right) \right], \end{aligned} \quad (15)$$

where  $D(\mathbf{p}_n | \mathbf{1})$  is the uniform Dirichlet distribution,  $\tilde{\boldsymbol{\alpha}}_n^v = \mathbf{y}_n + (\mathbf{1} - \mathbf{y}_n) \odot \boldsymbol{\alpha}_n^v$  is the Dirichlet parameters after removal of the non-misleading evidence from predicted parameters for the  $n$ -th instance, and  $\Gamma(\cdot)$  is the gamma function.

Therefore, given the Dirichlet distribution with parameter  $\boldsymbol{\alpha}_n^v$  for the  $n$ -th instance, the loss is:

$$\mathcal{L}_{acc}(\boldsymbol{\alpha}_n^v) = \mathcal{L}_{acc}(\boldsymbol{\alpha}_n^v) + \lambda_s \mathcal{L}_{KL}(\boldsymbol{\alpha}_n^v), \quad (16)$$

where  $\lambda_s = \min(1.0, s/\mathcal{T}) \in [0, 1]$  is the annealing coefficient, with  $t$  representing the current training epoch and  $\mathcal{T}$  the total number of annealing steps. This gradual increase in the influence of KL divergence in the loss function helps prevent premature convergence of misclassified instances towards a uniform distribution.

**Conflict-aware aggregation loss.** In the proposed S-MRF module, each graph node represents evidence derived from different views. To achieve conflict-aware MVC, it is essential to harmonize the evidence across these views. This is accomplished by a combined loss function that simultaneously maximizes the pairwise similarity between evidence and minimize the deviation of each evidence from their mean. Moreover, we introduce a regularizer to prevent node degradation and maintain diversity in the graph.

Specifically, to ensure global consistency and inter-view collaboration, we introduce a unified consistency loss  $\mathcal{L}_{\text{con}}$ , which combines both the similarity encouragement in terms of statistical property and regularization components:

$$\begin{aligned} \mathcal{L}_{\text{con}} &= - \frac{2}{V(V-1)} \sum_{i=1}^{V-1} \sum_{j=i+1}^V \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\| \|\mathbf{e}_j\|} \\ &\quad + \frac{1}{VN} \sum_{v=1}^V \sum_{n=1}^N \|\mathbf{e}_n^v - \boldsymbol{\mu}^v\|_2^2 - \sum_{v=1}^V \|\mathbf{E}^v\|_F^2, \end{aligned} \quad (17)$$

where  $\boldsymbol{\mu}^v$  represents the mean evidence across all samples within view  $v$ , while  $\mathbf{E}^v$  denotes the  $v$ -th view evidence.

To sum up, the overall loss function for a specific instance  $\{\mathbf{x}_n^v\}_{v=1}^V$  can be calculated as:

$$\mathcal{L} = \frac{1}{V} \sum_{v=1}^V \mathcal{L}_{acc}(\alpha_n^v) + \lambda_t \mathcal{L}_{con} \quad (18)$$

The model optimization process and extensive discussion can be found in **Supplementary Material**.

## Experiment

### Experimental setup

**Datasets.** The overview of the datasets used in our study is illustrated in Table 1, with more detailed descriptions provided in the **Supplementary Material**. In summary, we benchmarked our approach against state-of-the-art (SOTA) methods on eight multi-view datasets. The conflictive datasets were constructed by us based on these datasets, following the methodology outlined in (Xu et al. 2024). Each method was run 10 times, and we report the average performance along with the standard deviation.

Dataset	Size	Classes	Dimensionality
PIE	680	68	484 256 279
HandWritten	2000	10	240 76 216 47 64 6
Scene15	4485	15	20 59 40
Caltech101	8677	101	4096 4096
Animal	10158	50	4096 4096
ALOI	10800	100	77 13 64 125
CUB	11788	10	1024 300
NUS-WIDE-OBJECT	30000	31	129 74 145 226 65

Table 1: Overview of datasets used in experiments

**Compared methods.** In this study, we compared our approach against eight SOTA methods. Among these, TMC (Han et al. 2021), TMDL-OA (Liu et al. 2022), ECML (Xu et al. 2024), and RMVC (Yue et al. 2024) are EDL-based multi-view methods, while EDL (Sensoy, Kaplan, and Kandemir 2018) itself is a single-view method. Detailed descriptions of all the comparison methods can be found in the **Supplementary Material**.

### Quantative results

**Comparison with SOTA.** Tables 2 and 3 show the classification performance on normal and conflictive test sets. We can observe that: **(1)** On normal test sets, TUNED outperforms all other baselines on all datasets except for the HandWritten and Scene15 datasets. **(2)** When evaluating on conflictive test sets, the accuracy of all compared methods significantly decreases. However, TUNED achieves impressive results across all datasets, with particularly large improvements on the Scene15, Animal, and ALOI datasets (18%, 37.73%, and 15.4% respectively). Additionally, we notice that on the Animal dataset, the conflictive views even achieve better results than on the normal test sets. This might be due to the sensitivity of the parameters being further amplified by the reduced number of views, a phenomenon that has been validated in subsequent studies.

**Ablation study.** Table 4 presents the ablation study results of TUNED on the HandWritten and Scene datasets. The experimental results indicate that each component of our framework contributes positively to multi-view fusion. Moreover, compared to other fusion methods such as Arithmetic mean and DST, our MRF fusion demonstrates superior performance. We provide additional comparisons based on a broader set of metrics in the **Supplementary Material**.

### Qualitative results

**Visualization of S-MRF.** The Figure 3 illustrates the weight relationships between different views in the HandWritten dataset. The left and right panels show the conflict levels for normal and conflict instances, respectively. To create conflict, we modified the content of view 0, resulting in misalignment with the other views, and set the corresponding  $\tau$  to 0.7. The results demonstrate that, after training, the weights between the conflict view and the other views are reduced to zero. This clearly indicates that our fusion method effectively captures and mitigates the impact of conflict views on model integration.

**Robustness to conflict.** We evaluated TMDL-OA, ECML, and our TUNED model on the HandWritten dataset with varying numbers of noisy views in Figure 4. Conflict test sets were created by adding noise to single, paired, and multiple views among the six available. Results show that as more views are corrupted, accuracy generally decreases. However, our TUNED model consistently outperforms others, demonstrating lower variance and greater robustness, especially in scenarios with noise in single views, where ECML and TMDL-OA show significant instability.

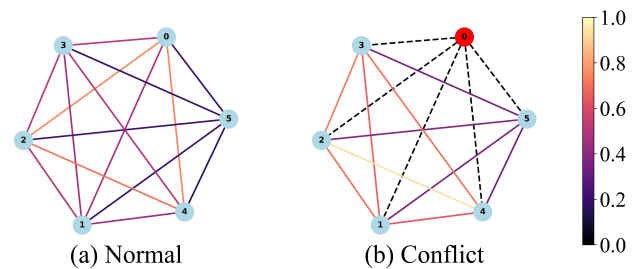


Figure 3: S-MRF weights for the HandWritten dataset. (a) Represents the weights under the normal setting, while (b) illustrates the weights under the conflict setting. And the red node indicates the view where conflict have been introduced.

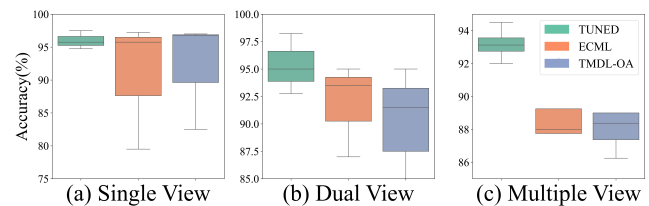


Figure 4: Robustness comparison: Impact of conflict views on model performance across different configurations.

Method	PIE	HandWritten	Scene15	Caltech101	CUB	Animal	ALOI	NUS
EDL	86.25±0.89	96.90±0.16	52.76±0.54	73.35±1.73	86.22±0.36	84.30±1.76	37.87±5.13	22.33±0.64
DCCAE	81.96±1.04	95.45±0.35	74.62±1.52	89.56±0.41	85.39±1.36	82.72±1.38	87.20±0.14	35.75±0.48
CPM-Nets	88.53±1.23	94.55±1.36	67.29±1.01	90.35±2.12	89.32±0.38	87.40±1.12	49.13±6.19	35.37±1.05
DUA-Nets	90.56±0.47	98.10±0.32	68.23±0.11	93.43±0.34	80.13±1.67	78.65±0.55	83.09±2.15	33.98±0.34
TMC	91.85±0.23	98.51±0.13	67.71±0.30	92.80±0.50	90.57±2.96	87.20±0.33	79.31±0.31	35.18±1.55
TMDL-OA	92.33±0.36	<u>99.25±0.45</u>	75.57±0.02	94.63±0.04	95.43±0.20	87.05±0.28	65.20±0.18	34.39±0.44
ECML	<u>94.71±0.02</u>	<b>99.40±0.00</b>	<b>76.19±0.12</b>	<u>95.36±0.38</u>	<u>98.50±2.75</u>	84.01±0.63	69.29±0.11	34.04 ± 0.27
RMVC	91.18±0.24	98.51±0.04	73.05±0.24	88.73±0.60	93.18±0.47	<u>87.67±0.17</u>	<u>87.38±0.67</u>	34.68±0.32
Ours	<b>96.83±0.01</b>	99.20±0.23	<u>75.89±0.70</u>	<b>97.38±0.45</b>	<b>99.00±0.33</b>	<b>89.52 ± 0.13</b>	<b>88.93±0.35</b>	<b>37.46±0.25</b>
△%	2.24	-0.2	-0.39	2.12	0.51	2.11	1.77	4.78

Table 2: Accuracy (%) on normal test sets. The best and the second best results are highlighted by **boldface** and underlined respectively. △% denotes the performance improvement of ours over the best baseline. NUS refers to “NUS-WIDE-OBJECT”.

Method	PIE	HandWritten	Scene15	Caltech-101	CUB	Animal	ALOI	NUS
EDL	21.76±0.67	57.25±0.49	14.28±0.24	55.74±0.12	53.75±0.42	30.71±0.27	25.05±3.10	18.07±0.28
DCCAE	26.89±1.10	82.85±0.38	25.97±2.86	60.90±2.32	63.57±1.28	64.30±2.11	75.12±0.43	32.12±0.52
CPM-Nets	53.19±1.17	83.34±1.07	29.63±1.12	66.54±2.89	68.82±0.17	64.83±0.35	36.29±5.02	29.20±0.81
DUA-Nets	56.45±1.75	87.16±0.34	26.18±1.31	75.19±2.34	60.53±1.17	62.46±1.12	69.07±2.50	31.82±0.43
TMC	61.65±1.03	92.76±0.15	42.27±1.61	90.16±2.50	73.37±2.16	64.85±1.19	76.68±0.32	33.76±2.16
TMDL-OA	68.16±0.34	93.05±0.05	48.42±1.02	90.63±2.05	74.43±0.26	64.62±0.15	62.90±0.12	32.44±0.26
ECML	<u>84.00±0.14</u>	94.40±0.05	<u>56.97±0.52</u>	<u>92.36±1.48</u>	<u>76.50±1.15</u>	62.67±0.81	64.91 ± 0.20	31.19 ± 0.22
RMVC	<u>76.47±3.43</u>	<u>94.75±0.75</u>	<u>49.83±2.23</u>	<u>80.56±0.71</u>	<u>72.78±0.42</u>	<u>66.00±0.59</u>	52.67±1.97	24.62±3.19
Ours	<b>86.02±0.19</b>	<b>96.75± 0.55</b>	<b>67.22± 0.58</b>	<b>93.22±0.41</b>	<b>76.67±0.38</b>	<b>90.90± 0.18</b>	<b>88.49±0.29</b>	<b>34.09±0.14</b>
△%	2.40	2.11	18.00	0.93	2.22	37.73	15.40	0.98

Table 3: Accuracy (%) on conflictive test sets. The tags setting is the same as that in Table 2.

Method	HandWritten		Scene15	
	Normal	Conflict	Normal	Conflict
w/o GCN	95.75	95.75	69.23	64.21
w/o $e_n^{\text{cons}}$	96.50	94.25	70.79	64.88
w/o $\mathcal{L}_{\text{con}}$	96.50	94.50	70.79	65.44
w/ Average	96.25	96.00	70.90	65.33
w/ DST	95.75	91.75	70.56	66.67
w/ S-MRF	<b>99.20</b>	<b>96.75</b>	<b>75.89</b>	<b>67.22</b>

Table 4: Ablation study: Accuracy (%) on HandWritten and Scene15 datasets under normal and conflict conditions.

**T-SNE visualization.** We visualized the Evidence from three views and their fused representation on the Scene dataset after model training in Figure 5. The results demonstrate that MRF effectively captures inter-view relationships, producing a fused Evidence representation with more distinct class separations, thereby enhancing the discriminative power between categories.

## Conclusion

In this study, we introduced the TUNED model to address the challenges of multi-view classification, particularly in scenarios with conflicting views and high uncertainty. Our approach successfully integrates local and global feature-neighborhood structures, significantly enhancing robustness and driving improvements in classification performance. A limitation is its untested application in incomplete multi-view data scenarios, where missing views may degrade the structure. Future work will extend the approach to robustly handle incomplete multi-view classification.

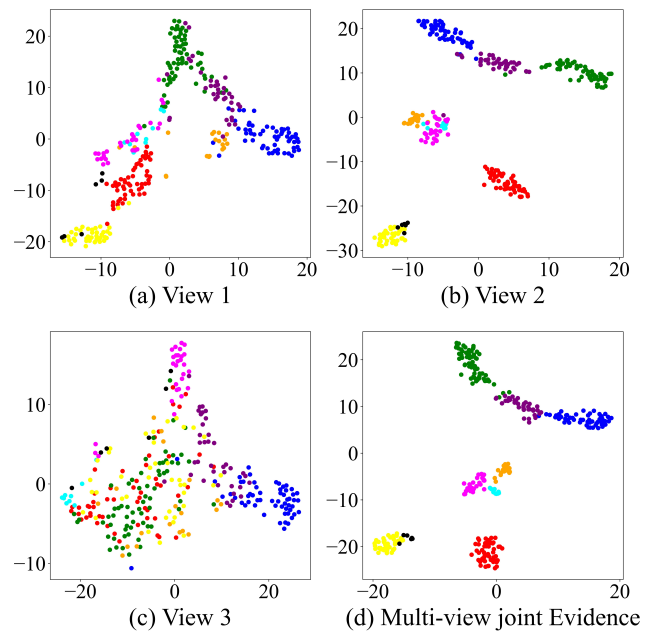


Figure 5: T-SNE visualization of Evidence. (a)-(c) show the t-SNE results for Evidence from three individual views. (d) illustrates the fused Evidence after t-SNE, highlighting the more cohesive data representation achieved through fusion.

## References

- Bao, W.; Yu, Q.; and Kong, Y. 2021. Evidential Deep Learning for Open Set Action Recognition. *arXiv:2107.10161*.
- Chen, H.; Huang, H.; Dong, J.; Zheng, M.; and Shao, D. 2024a. FineCLIPER: Multi-modal Fine-grained CLIP for Dynamic Facial Expression Recognition with AdaptERs. *arXiv preprint arXiv:2407.02157*.
- Chen, H.; Huang, Y.; Huang, H.; Ge, X.; and Shao, D. 2024b. GaussianVTON: 3D Human Virtual Try-ON via Multi-Stage Gaussian Splatting Editing with Image Prompting. *arXiv preprint arXiv:2405.07472*.
- Chen, J.; Ma, K.; Huang, H.; Shen, J.; Fang, H.; Zang, X.; Ban, C.; He, Z.; Sun, H.; and Kang, Y. 2024c. BoViLA: Bootstrapping Video-Language Alignment via LLM-Based Self-Questioning and Answering. *arXiv preprint arXiv:2410.02768*.
- Chen, M.; Gao, J.; and Xu, C. 2023. Uncertainty-Aware Dual-Evidential Learning for Weakly-Supervised Temporal Action Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 15896–15911.
- Dempster, A. P. 1968. A generalization of Bayesian inference. *J R Stat Soc B*, 30(2): 205–232.
- Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 1050–1059. PMLR.
- Gao, J.; Chen, M.; and Xu, C. 2023. Vectorized Evidential Learning for Weakly-Supervised Temporal Action Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 15949–15963.
- Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2021. Trusted multi-view classification. In *ICLR*.
- Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2022. Trusted Multi-View Classification with Dynamic Evidential Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Hou, D.; Cong, Y.; Sun, G.; Dong, J.; Li, J.; and Li, K. 2020. Fast multi-view outlier detection via deep encoder. *IEEE Transactions on Big Data*, 8(4): 1047–1058.
- Huang, H.; Liu, Z.; Han, X.; Yang, X.; and Liu, L. 2023a. A belief logarithmic similarity measure based on dempster-shafer theory and its application in multi-source data fusion. *J. Intell. Fuzzy Syst.*, (Preprint): 1–13.
- Huang, H.; Liu, Z.; Letchmunan, S.; Deveci, M.; Lin, M.; and Wang, W. 2024a. Evidential deep partial multi-view classification with discount fusion. *arXiv preprint arXiv:2408.13123*.
- Huang, H.; Qiao, X.; Chen, Z.; Chen, H.; Li, B.; Sun, Z.; Chen, M.; and Li, X. 2024b. CREST: Cross-modal Resonance through Evidential Deep Learning for Enhanced Zero-Shot Learning. *arXiv preprint arXiv:2404.09640*.
- Huang, Z.; Hu, P.; Zhou, J. T.; Lv, J.; and Peng, X. 2020. Partially view-aligned clustering. *Advances in Neural Information Processing Systems*, 33: 2892–2902.
- Huang, Z.; Ren, Y.; Pu, X.; Huang, S.; Xu, Z.; and He, L. 2023b. Self-Supervised Graph Attention Networks for Deep Weighted Multi-View Clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7): 7936–7943.
- Jøsang, A. 2016. *Subjective Logic*, volume 3. Springer.
- Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *NIPS*, 30.
- Lampert, C. H.; and Krömer, O. 2010. Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning. In *Computer Vision—ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5–11, 2010, Proceedings, Part II 11*, 566–579. Springer.
- Li, X. 2022. Positive-incentive noise. *IEEE Transactions on Neural Networks and Learning Systems*.
- Liu, W.; Chen, Y.; Yue, X.; Zhang, C.; and Xie, S. 2023. Safe Multi-View Deep Classification. In *AAAI*, 8870–8878.
- Liu, W.; Yue, X.; Chen, Y.; and Denoeux, T. 2022. Trusted Multi-View Deep Learning with Opinion Aggregation. *AAAI*, 36(7): 7585–7593.
- Liu, Z.; Huang, H.; and Letchmunan, S. 2023. Adaptive weighted multi-view evidential clustering. In *International Conference on Artificial Neural Networks*, 265–277. Springer.
- Liu, Z.; Huang, H.; Letchmunan, S.; and Deveci, M. 2024. Adaptive weighted multi-view evidential clustering with feature preference. *Knowl.-Based Syst.*, 294: 111770.
- Lyzhov, A.; Molchanova, Y.; Ashukha, A.; Molchanov, D.; and Vetrov, D. 2020. Greedy policy search: A simple baseline for learnable test-time augmentation. In *UAI*, 1308–1317. PMLR.
- Ma, K.; Huang, H.; Chen, J.; Chen, H.; Ji, P.; Zang, X.; Fang, H.; Ban, C.; Sun, H.; Chen, M.; et al. 2024. Beyond Uncertainty: Evidential Deep Learning for Robust Video Temporal Grounding. *arXiv preprint arXiv:2408.16272*.
- Marcos Alvarez, A.; Yamada, M.; Kimura, A.; and Iwata, T. 2013. Clustering-based anomaly detection in multi-view data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 1545–1548.
- Nie, F.; Wang, X.; and Huang, H. 2014. Clustering and projected clustering with adaptive neighbors. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 977–986.
- Qin, Y.; Chen, Y.; Peng, D.; Peng, X.; Zhou, J. T.; and Hu, P. 2024. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27197–27206.
- Qin, Y.; Peng, D.; Peng, X.; Wang, X.; and Hu, P. 2022. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4948–4956.
- Sensoy, M.; Kaplan, L. M.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31: 3183–3193.

Shao, Z.; Dou, W.; and Pan, Y. 2024. Dual-level Deep Evidential Fusion: Integrating multimodal information for enhanced reliable decision-making in deep learning. *Information Fusion*, 103: 102113.

Wang, X.; Peng, D.; Hu, P.; and Sang, Y. 2019. Adversarial correlated autoencoder for unsupervised multi-view representation learning. *Knowledge-Based Systems*, 168: 109–120.

Wen, J.; Liu, C.; Deng, S.; Liu, Y.; Fei, L.; Yan, K.; and Xu, Y. 2023a. Deep Double Incomplete Multi-View Multi-Label Learning With Incomplete Labels and Missing Views. *IEEE Transactions on Neural Networks and Learning Systems*, 1–13.

Wen, J.; Liu, C.; Xu, G.; Wu, Z.; Huang, C.; Fei, L.; and Xu, Y. 2023b. Highly Confident Local Structure Based Consensus Graph Learning for Incomplete Multi-View Clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15712–15721.

Wen, Y.; Wang, S.; Liao, Q.; Liang, W.; Liang, K.; Wan, X.; and Liu, X. 2023c. Unpaired Multi-View Graph Clustering With Cross-View Structure Matching. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15.

Xiao, F. 2019. Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy. *Information Fusion*, 46: 23–32.

Xu, C.; Si, J.; Guan, Z.; Zhao, W.; Wu, Y.; and Gao, X. 2024. Reliable Conflictive Multi-View Learning. arXiv:2402.16897.

Yang, M.; Li, Y.; Huang, Z.; Liu, Z.; Hu, P.; and Peng, X. 2021. Partially View-Aligned Representation Learning With Noise-Robust Contrastive Loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1134–1143.

Yu, Y.; Deng, D.; Liu, F.; Jin, Y.; Dou, Q.; Chen, G.; and Heng, P.-A. 2023. Adaptive Negative Evidential Deep Learning for Open-set Semi-supervised Learning. arXiv:2303.12091.

Yue, X.; Dong, Z.; Chen, Y.; and Xie, S. 2024. Evidential dissonance measure in robust multi-view classification to resist adversarial attack. *Information Fusion*, 102605.

Zhang, P.; Wang, S.; Li, L.; Zhang, C.; Liu, X.; Zhu, E.; Liu, Z.; Zhou, L.; and Luo, L. 2023. Let the Data Choose: Flexible and Diverse Anchor Graph Fusion for Scalable Multi-View Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11262–11269.

Zhang, T.; Liu, X.; Gong, L.; Wang, S.; Niu, X.; and Shen, L. 2021. Late fusion multiple kernel clustering with local kernel alignment maximization. *IEEE Transactions on Multimedia*.

Zhao, H.; Liu, H.; Ding, Z.; and Fu, Y. 2017. Consensus regularized multi-view outlier detection. *IEEE Transactions on Image Processing*, 27(1): 236–248.