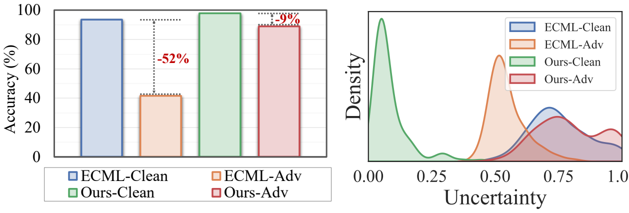
Reliable Disentanglement Multi-view Learning Against View Adversarial Attacks

Xuyang Wang[1] , **Siyuan Duan**[1] , **Qizhi Li**[1] , **Guiduo Duan**[2] , **Yuan Sun**[1*] and **Dezhong Peng**[1]
[1]College of Computer Science, Sichuan University
[2]Laboratory of Intelligent Collaborative Computing, University of Electronic Science and Technology of China
xywang@stu.scu.edu.cn, {ddzz122773315, mrqz945, sunyuan_work}@163.com,
guiduo.duan@uestc.edu.cn, pengdz@scu.edu.cn

Abstract

Recently, trustworthy multi-view learning has attracted extensive attention because evidence learning can provide reliable uncertainty estimation to enhance the credibility of multi-view predictions. Existing trusted multi-view learning methods implicitly assume that multi-view data is secure. In practice, however, in safety-sensitive applications such as autonomous driving and security monitoring, multi-view data often faces threats from adversarial perturbations, thereby deceiving or disrupting multi-view learning models. This inevitably leads to the adversarial unreliability problem (AUP) in trusted multi-view learning. To overcome this tricky problem, we propose a novel multi-view learning framework, namely Reliable Disentanglement Multi-view Learning (RDML). Specifically, we first propose evidential disentanglement learning to decompose each view into clean and adversarial parts under the guidance of corresponding evidences, which is extracted by a pretrained evidence extractor. Then, we employ the feature recalibration module to mitigate the negative impact of adversarial perturbations and extract potential informative features from them. Finally, to further ignore the irreparable adversarial interferences, a view-level evidential attention mechanism is designed. Extensive experiments on multi-view classification tasks with adversarial attacks show that our RDML outperforms the state-of-the-art multi-view learning methods by a relatively large margin. Our code is available at https://github.com/Willy1005/2025-IJCAI-RDML.

1 Introduction



(a) Classification accuracy (b) Uncertainty density

Figure 1: We conduct experiments on the PIE dataset in clean and adversarial settings, and show a toy example of AUP. Note that only one view is randomly attacked via PGD. (a) shows the classification accuracy of the ECML method and our RDML. (b) presents the estimated uncertainties of these methods.

In practical scenarios, an object can often be described by multiple views of different feature types and modalities, which leads to a growing interest in multi-view learning. Thanks to the power of deep learning, deep multi-view learning has exhibited remarkable advantages by integrating and mining both valuable complementary and consistency information of multi-views [Qin *et al.*, 2024; Zhang *et al.*, 2020; Sun *et al.*, 2024]. Thus, in recent years, multi-view learning has attracted widespread attention [Zhang *et al.*, 2023]. For example, predictive dynamic fusion (PDF) [Cao *et al.*, 2024] proposes an intuitive and rigorous multimodal fusion paradigm from the perspective of generalization error.

Although these above methods have achieved pleasing performance, their results could be uncertain and unreliable due to the attribute differences and heterogeneity of multi-view data. This greatly limits the application of multi-view learning in various fields, especially medical diagnosis or autonomous driving. To this end, a new trusted learning paradigm for multi-view classification is proposed to enhance trusted decisions. For instance, Trusted Multi-view Classification (TMC) [Han *et al.*, 2022b] introduces the evidence theory to construct the Dirichlet distribution, thereby providing uncertainty estimation for multi-view decisions to enhance reliability. To make reliable decisions under noise labels, TMNR [Xu *et al.*, 2024b] proposes trusted multi-view noise refining to overcome the negative effects of noisy labels.

Almost all existing trusted multi-view learning methods implicitly assume that multi-view data is secure. In practice, however, in safety-critical applications such as autonomous driving and security monitoring, multi-view data could be susceptible to adversarial attacks, which can deceive or disrupt multi-view learning models. This vulnerability inevitably leads to the adversarial unreliability problem (AUP) in trusted multi-view learning. As shown in Fig.1, adversarial

[*]Corresponding author

attacks (such as projected gradient descent attack [Madry *et al.*, 2018]) are imposed on a view of multi-view data. From the figure, we can observe that, after being subjected to adversarial attacks, even if only one view is attacked, the state-of-the-art evidence-based method (i.e., ECML) still shows a significant decline in classification performance. Worse still, instead of increasing with the substantial performance decline, the estimated uncertainties are significantly lower than those in the clean setting. This indicates that the evidence-based uncertainty estimation mechanism fails under adversarial perturbations, thereby leading to the AUP.

To overcome the above AUP, we propose a novel multi-view learning framework, namely Reliable Disentanglement Multi-view Learning (RDML). In the first stage, we present a perturbation-insensitive pretraining scheme to train an evidence extractor, thereby providing reliable category-level evidence and uncertainty estimation in subsequent stages. In the second stage, as shown in Fig.2, our method mainly consists of three key modules: the evidential disentanglement learning module, the feature recalibration module, and the evidential attention module. To be specific, RDML first proposes evidential disentanglement learning to decompose each view into clean and adversarial features under the guidance of pretrained corresponding evidence. Then, to prevent the negative effect of the adversarial features, we propose feature recalibration to recalibrate these feature units for additional discriminative information, thereby obtaining more robust features. Finally, to reduce the interference of stubborn adversarial features that are difficult to calibrate, we design a view-level evidential attention mechanism to enhance the robustness against adversarial features. The main contributions of this paper are as follows.

- This paper studies a less-touched adversarial unreliability problem (AUP) in trusted multi-view learning and proposes a new Reliable Disentanglement Multi-view Learning (RDML) framework against view adversarial attacks. To the best of our knowledge, for the first time, we address the AUP caused by view adversarial attacks.
- We propose evidential disentanglement learning to guide the stripping of adversarial information from multi-view representations using a pretrained evidence model. To mitigate the interference of perturbations, we propose feature recalibration to rectify the weak adversarial features, and further present evidence attention to deal with the stubborn adversarial features.
- We conduct extensive experiments on six multi-view datasets to verify the effectiveness and robustness of our RDML under both adversarial and clean conditions.

2 Related Work

2.1 Trusted Multi-view Learning

Deep multi-view learning utilizes view-specific deep representations to enhance the integration and understanding of multi-source information [Geng *et al.*, 2021; Peng *et al.*, 2019; Xu *et al.*, 2024c]. Han et al. point out that existing studies have placed too much emphasis on improving the performance of deep multi-view learning methods in various scenarios while neglecting to enhance the reliability of multi-view decision [Han *et al.*, 2020]. They then propose the Trusted Multi-view Classification method. TMC transforms traditional classification networks into evidential neural networks (by replacing the final softmax function of the classification networks with Relu function to ensure that each output value is non-negative) and uses the extracted evidence representations to model the Dirichlet distribution [Sensoy *et al.*, 2018; Jsang, 2018]. The Dirichlet distribution can output classification category probability and uncertainty, thereby achieving reliable multi-view decision. Finally, TMC introduces Dempster rule to fuse multi-view opinions.

Xu et al. introduce evidence learning into conflictive multi-view learning, extracting view-specific opinion through the parameterized Dirichlet distribution [Xu *et al.*, 2024a]. And a conflictive opinion aggregation method is designed for multi-view fusion. Liu et al. propose an opinion fusion method based on evidence accumulation, in which the evidence representations of different views are accumulated to obtain the overall opinion [Liu *et al.*, 2022]. Yue et al. discover the vulnerability of trustworthy multi-view approaches to adversarial examples and attributed this vulnerability to the difficulty of accurately assessing the quality of adversarial examples [Yue *et al.*, 2025]. In this paper, we address the adversarial unreliability problem from a more fundamental perspective. That is, we first enhance the adversarial insensitivity of the evidence neural network (perturbation-insensitive pretraining), and subsequently minimize the interference and harm that adversarial perturbations inflict on multi-view fusion and decision to the greatest extent possible (evidence-based disentanglement, feature recalibration and view-level evidential attention).

2.2 Deep Adversarial Defense

Improving the robustness of deep neural networks against adversarial perturbations has long been a goal in deep learning community [Long *et al.*, 2022; Wang *et al.*, 2022; Kurakin *et al.*, 2018]. Since there are different defense strategies for different types of attacks, here we mainly focus on deep defense methods against white-box attacks. In recent years, adversarial training has been widely proven to be effective in enhancing the adversarial robustness of neural networks [Goodfellow *et al.*, 2015; Ilyas *et al.*, 2019]. By using both clean samples and adversarial samples as training data, the insensitivity of neural networks to adversarial perturbations can be enhanced.

Moreover, since disentanglement learning is good at separating information and features, it is naturally suitable for dealing with adversarial samples [Kim *et al.*, 2023; Liu *et al.*, 2024; Zhang *et al.*, 2024]. Typically, disentanglement learning decomposes perturbed samples into clean features and adversarial features to mitigate the interference caused by adversarial perturbations. However, existing disentanglement learning methods lack the support of effective clues and trusted guidance, and thus the adversarial robustness of models is limited. In this paper, we propose an evidence-based disentanglement method to resist adversarial perturbations. The disentanglement process is guided by a pretrained evi-

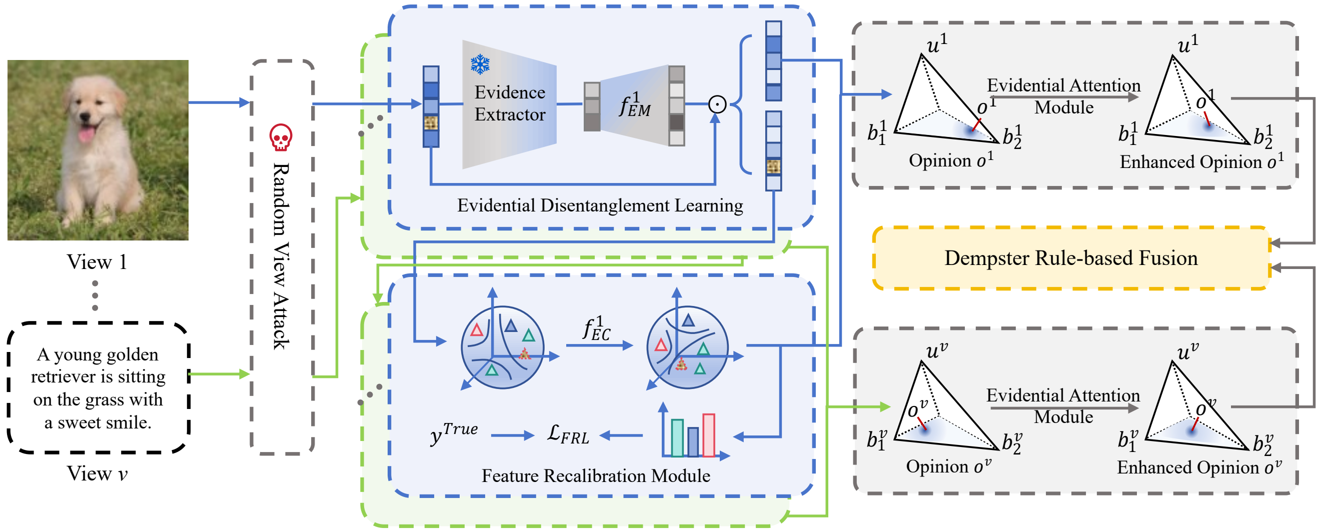


Figure 2: The framework of the proposed RDML, which is consists of four modules. First, evidential disentanglement learning uses the pretrained evidence extractor to conduct a robustness analysis of features under random view attacks, and generate a robustness mask to decouple clean and adversarial features. Next, the feature recalibration module will rectify the adversarial features. For features that are difficult to repair, RDML will generate evidence attention under the guidance of the evidence extractor to further mitigate the interference of adversarial features. Finally, RDML introduces the Dempster Rule-based Fusion method for opinion aggregation.

dence extractor, so it can improve the reliability and adversarial robustness of the evidential multi-view method.

3 Method

3.1 Problem Definition

Suppose that there is a multi-view dataset D_N^V with N instances and V views, and $x_n^v \in \mathbb{R}^{d_v}$ $(n = 1, 2, ..., N)$ is an instance from D_N^V. d_v is the dimension of the v-th view. The corresponding class labels are $\{\mathbf{y}_n\}_{n=1}^N$. The number of classes is K. The random view adversarial attacks are conducted to each multi-view instance by the following formula,

$$\widehat{x}_n^v = \text{AdvAttack}(x_n^v), \tag{1}$$

where d_v is the dimension of the v-th view. $\widehat{x}_n^v$ means the adversarial version of x_n^v. Our goal is to learn a robust evidential model against view adversarial attacks.

3.2 Overview

When facing adversarial attacks, existing trusted multi-view learning methods can easily lead to the adversarial unreliability problem (AUP) in trusted multi-view learning systems, thereby weakening the performance of multi-view learning models. To address the above problem, in this paper, we propose Reliable Disentanglement Multi-view Learning (RDML) against view adversarial attacks, which consist of two stages. In the first stage, we propose a perturbation-insensitive pretraining scheme to enhance the stability and adversarial insensitivity of evidential neural networks. To be specific, this approach introduces random view attacks into the pretraining of multi-view evidential networks, thereby making the evidence extractor provide the support of reliable category-level evidence and uncertainty estimation. The objective function could be formulated as follows:

$$\mathcal{L}_{PT} = \mathcal{L}_{ECL}(\alpha_n) + \sum_{v=1}^{V} \mathcal{L}_{ECL}(\alpha_n^v) + \mathcal{L}_{ACL}, \tag{2}$$

where the $\mathcal{L}_{ECL}(\alpha_n)$ is the evidential classification loss; α_n is the Dirichlet parameter; $\mathcal{L}_{ACL}$ denotes the adversarial consistency loss.

In the second stage, RDML first proposes evidential disentanglement learning. To be specific, we use the pretrained evidence extractor to analyze the features to be decoupled and map the extracted category-level evidence into a robustness-aware soft mask. The higher the score of the mask for a feature, the more likely the feature is to be a robust (or clean) feature, and vice versa, it is more likely to be an adversarial feature. Afterward, we utilize the mask to separate clean and adversarial features. In addition, we believe that some weak adversarial features could easily be converted into clean features. To this end, we propose feature recalibration to correct these adversarial features into clean features. For the remaining part of stubborn adversarial features, we design a view-level evidential attention mechanism to reduce the interference of these adversarial features that are difficult to correct, thereby enhancing the robustness against adversarial features. The objective function could be expressed as

$$\begin{aligned} \mathcal{L}_{\mathcal{T}} =& \mathcal{L}_{ECL}(\alpha_n) + \sum_{v=1}^{V} \mathcal{L}_{ECL}(\alpha_n^v) + \mathcal{L}_{ACL} \\ &+ \gamma_1 \mathcal{L}_{EDL} + \gamma_2 \mathcal{L}_{FRL}, \end{aligned} \tag{3}$$

where $\mathcal{L}_{EDL}$ and $\mathcal{L}_{FRL}$ represent evidential disentanglement loss and feature recalibration loss, respectively; γ_1 and γ_2 are two balancing parameters.

3.3 Perturbation-insensitive Pretraining

Although evidential neural networks can provide corresponding uncertainties to enhance the ability of trusted decisions, they could become unstable and even ineffective after being subjected to adversarial attacks. To overcome this issue, we propose a perturbation-insensitive pretraining strategy that incorporates adversarial samples to enhance the adversarial insensitivity of the evidence extractor $E_{pt}(\cdot) = \{E_{pt}^v(\cdot)\}_{v=1}^V$, thereby providing robust evidence support. To be specific, for adversarial multi-view data, to enhance the robustness of the evidence extractor for adversarial attacks, we then mix the adversarial multi-view samples with clean samples and utilize the evidence extractor for the mixed samples x_{mix} to extract the corresponding classification evidence e, i.e.,

$$e_k^v = E_{pt}^v(x_{mix}^v), k = 1, 2, ..., K, \tag{4}$$

where K denotes the number of classes. Afterward, we obtain the parameter $\alpha_k^v = e_k^v + 1$ required for modeling the Dirichlet distribution based on the evidence. The view-specific opinion $o^v = (b^v, u^v)$ are also obtained according to following formula,

$$b_k^v = \frac{e_k^v}{S^v}, u^v = \frac{K}{S^v}, \tag{5}$$

where $b \geq 0$ and $u \geq 0$ denote the belief mass and uncertainty ($\sum_{k=1}^K b_k^v + u^v = 1$), and $S^v = \sum_{k=1}^K \alpha_k^v$ represents the Dirichlet strength. Given the S^v and α_k^v, the probability of the v-th view for the k-th class is $p_k^v = \frac{\alpha_k^v}{S^v}$. After that, Dempster rule is used to combine multi-view opinions, and the joint opinion is $o = o^1 \otimes o^2 \otimes ... \otimes o^v$, where $\otimes$ is the Dempster rule based fusion operation.

To optimize the evidence extractor, following [Han *et al.*, 2020], we use the evidential classification loss $\mathcal{L}_{ECL}(\alpha_n)$, i.e.,

$$\mathcal{L}_{ECL}(\alpha_n) = \mathcal{L}_{ECE} + KL[D(p_n \mid \tilde{\alpha}_n) \parallel D(p_n \mid 1)], \tag{6}$$

$$\mathcal{L}_{ECE}(\alpha_n) = \sum_{k=1}^K y_{nk}(\psi(S_n) - \psi(\alpha_n)), \tag{7}$$

where $\psi(\cdot)$ denotes the digamma function; $\tilde{\alpha}_n = y_n + (1 - y_n) \odot \alpha_n$ is the adapted Dirichlet parameter avoiding penalizing the evidence of the correct category to zero. $\mathcal{L}_{ECE}(\alpha_n)$ represents the evidential cross-entropy loss, which requires the model to extract more evidence for the correct category compared to other categories; and the Kullback-Leibler (KL) divergence restricts the model to extract as little evidence as possible from the incorrect categories.

Besides, since the instances in AMVL are subjected to random view attacks, we hope to enhance model robustness against adversarial attacks by constraining the differences among the predicted probability distributions of different views. Therefore, we construct the following adversarial consistency loss $\mathcal{L}_{ACL}$, i.e.,

$$\mathcal{L}_{ACL} = \frac{1}{V-1}\sum_{v_1=1}^{V}(\sum_{v_2 \neq v_1}^{V} \frac{\sum_{k=1}^K |p_k^{v_1} - p_k^{v_2}|}{2}), \tag{8}$$

Therefore, the well-trained evidence extractor will guarantee robust and stable evidence in subsequent steps with frozen parameters.

3.4 Evidential Disentanglement Learning

Although existing disentanglement learning methods can be used for decomposing adversarial and clean features by various meticulously designed losses, the lack of support for effective clues and trusted guidance hinders their performance. To improve the credibility of separating adversarial features, we introduce the evidential learning theory to naturally provide evidence and uncertainty for the decision process. This characteristic can significantly promote reliable feature decomposition. To this end, we design an evidence disentanglement learning module.

In the training stage, given a clean multi-view instance $x = \{x^v \in \mathbb{R}^{d_v}\}_{v=1}^V$, the random view attack is conducted on it, and its adversarial version is $\widehat{x}$. Different from existing disentanglement learning methods, we leverage the pre-trained evidence extractor $E_{\text{pt}}(\cdot)$ to extract adversarially-insensitive evidence from the adversarial samples. Due to strong adversarial insensitivity, pretrained evidence extractor $E_{pt}(\cdot)$ can extract effective evidence under adversarial perturbation, i.e.,

$$em^v = E_{pt}^v(\widehat{x}^v), \tag{9}$$

where $em^v \in \mathbb{R}^K$ is the evidential map, which implies the category-aware evidence of the v-th view of adversarial sample $\widehat{x}^v$. Since em^v is category-level evidence, we construct a multi-layer perceptron (MLP) based evidence mapping layer $f_{EM,1}(\cdot) = \{f_{EM,1}^v(\cdot)\}_{v=1}^V$ to map the category-level evidence em^v into a feature-level robustness map $rm^v = f_{EM,1}^v(em^v)$ ($rm^v \in \mathbb{R}^{d_v}$), which indicates the amount of evidence for features in each dimension. Features containing more evidence are regarded as clean features, while those containing less are considered adversarial features. Subsequently, in order to facilitate the decomposition of these two types of features, we introduce Gumbel softmax [Jang *et al.*, 2022] to convert the robustness map into a soft mask score m^v, i.e.,

$$m^v = \frac{e^{(\log(\sigma(rm^v)) + q_1)/\mu}}{e^{(\log(\sigma(rm^v)) + q_1)/\mu} + e^{(\log(1-\sigma(rm^v)) + q_2)/\mu}}, \tag{10}$$

where each value is a non-negative value less than 1. $\sigma(\cdot)$ represents a Sigmoid function for normalization; q_1 and q_2 are two instances sampled from Gumbel distribution (given $u \sim$ Uniform(0, 1), $q = -\log(-\log(u))$); μ is a temperature coefficient. $m^v \in \mathbb{R}^{d_v}$ is a feature-level mask score and each dimension of the feature has a score ranging from 0 to 1, a higher score indicates a higher probability of being a clean feature.

Therefore, we can decompose clean and adversarial features via the following simple feature-level multiplication,

$$h_c^v = x^v \odot m^v, \tag{11}$$

$$h_a^v = x^v \odot (1 - m^v), \tag{12}$$

where h_c^v and h_a^v are the clean feature of x^v and the adversarial one, respectively.

To optimize our evidential disentanglement learning, an evidence disentanglement loss is designed. Evidential disentanglement learning utilizes the evidence output by a pretrained evidence extractor to generate a soft robustness mask. Therefore, we expect that the distribution of the decoupled clean features is as close as possible to the real distribution, while the distribution of adversarial features is in contrast. Given clean and adversarial feature of the n-th instance $h_{c,n}$, $h_{a,n}$, the evidence disentanglement loss can be written as

$$\mathcal{L}_{EDL} = -\sum_{v=1}^{V} (y_n \odot \log(p_{c,n}^v) + \widehat{y}_n \odot \log(p_{a,n}^v)), \quad (13)$$

where y_n is the ground truth label of h_n; $\widehat{y}_n$ is the label of a wrong class for h_n; $p_{c,n}^v = f_{EC}^v(h_{c,n}^v)$, $p_{a,n}^v = f_{EC}^v(h_{a,n}^v)$ are the classification probabilities of $h_{c,n}^v$ and $h_{a,n}^v$; $f_{EC}(\cdot) = \{f_{EC}^v(\cdot)\}_{v=1}^V$ is a group of evidential classifiers where an activation function like Relu is added after each classifier. Note that $f_{EC}(\cdot)$ only participates in the training phase and is not utilized in the pretraining and testing phase.

3.5 Feature Recalibration

For the adversarial features, we believe that a part of them can be transformed from the clean features with relatively poor robustness, and this part of features also easily has the potential to be recovered back into clean features. Therefore, we construct an MLP based feature recalibration layer $f_{FC}(\cdot) = \{f_{FC}^v(\cdot)\}_{v=1}^V$ to rectify weak adversarial features to clean and informative features,

$$h_{cr}^v = f_{FC}^v(h_a^v), \quad (14)$$

where h_{cr}^v represents the corrected feature of $\widehat{x}^v$. Then we can get the final feature $h_f^v = h_c^v + h_{cr}^v$. For the feature recalibration module, we expect that the corrected representations can provide as much informative and valuable features as possible for classification. Therefore, the predicted probability distribution of the corrected representations is required to be as close as possible to the ground truth distribution. Thus, we can have the following loss, i.e.,

$$\mathcal{L}_{FRL} = -\sum_{v=1}^{V} y_n \odot \log(f_{EC}^v(h_{cr,n}^v)). \quad (15)$$

3.6 Evidential Attention

Though decoupling adversarial features and repairing weak adversarial features can alleviate the interference of adversarial perturbations on evidential neural networks to some extent, the impairment brought about by stubborn adversarial features remains difficult to mitigate effectively. Thus, we propose a view-level evidential attention mechanism that generates the evidential attention score by conducting a robust analysis of view features using the pre-trained $E_{pt}(\cdot)$. This mechanism guides the model to focus on the informative clean features and ignore the interference of strong adversarial features that are difficult to utilize. Specifically, we utilize the pretrained evidence extractor $E_{pt}(\cdot)$ to conduct evidence analysis on the feature h_f. The extracted evidence is transformed into evidential attention scores via a softmax function. Benefiting from the knowledge of the pre-trained evidence extractor, these attention scores imply the robustness of each dimension of h_f, distinguishing clean, weak adversarial, and hard adversarial features in the form of scores. Since the evidence extractor outputs category-level scores, we map them into feature-level attentions through the evidence mapping layer $f_{EM,2}(\cdot) = \{f_{EM,2}^v(\cdot)\}_{v=1}^V$, i.e.,

$$att_i^v = \frac{e^{E_{pt}^v(h_{f,i}^v)}}{\sum_{j=1}^{K} e^{E_{pt}^v(h_{f,j}^v)}}, \quad (16)$$

$$att^v = f_{EM,2}^v(att^v), \quad (17)$$

$h_{f,i}^v$ denotes the i-th element of h_f^v; $att^v \in \mathbb{R}^{d_v}$ is the view-level evidential attention for h_f^v. The augmented feature is obtained via $h_{aug}^v = h_f^v \odot att^v$.

3.7 Trusted Multi-view Fusion

Followed by [Han *et al.*, 2020], Dempster combination rule is introduced for multi-view fusion. Given two augmented features of two views h_{aug}^1, h_{aug}^2, the corresponding evidences $e^1 = E_c^1(h_{aug}^1)$, $e^2 = E_c^2(h_{aug}^2)$ can be extracted via a evidence extractor $E_c(\cdot) = \{E_c^v(\cdot)\}_{v=1}^V$. It is worth mentioning that the parameters of $E_c(\cdot)$ are copied from the pretrained $E_{pt}(\cdot)$, so as to improve the overall training efficiency. Then two opinions $o^1 = (b^1, u^1)$ and $o^2 = (b^2, u^2)$ are constructed via Eq. (5). After that, we have joint opinion $o = o^1 \otimes o^2 = (b, u)$ where

$$b_k = \frac{b_k^1 b_k^2 + b_k^2 u^1 + b_k^1 u^2}{1 - M}, u = \frac{u^1 u^2}{1 - M}. \quad (18)$$

$M = \sum_{i \neq j} b_i^1 b_j^2$ represents the difference between two opinions. And $\frac{1}{1-M}$ is used for normalization. According to the above combination pattern, we have the joint multi-view opinion $o = o^1 \otimes o^2 \otimes ... \otimes o^v = (b, u)$. Then the joint evidence $e_k = b_k \times S$, Dirichlet parameter $\alpha_k = e_k + 1$ and Dirichlet strength $S = \frac{K}{u}$ are obtained based on Eq. (5).

4 Experiments

4.1 Datasets and Competitors

To verify the effectiveness and robustness of our method, we conduct extensive experiments on six multi-view datasets, including PIE [Gross *et al.*, 2010], Scene [Fei-Fei and Perona, 2005], Leaves [Cope *et al.*, 2013], NUS-WIDE [Chua *et al.*, 2009], MSRC [Xu *et al.*, 2016], and Fashion [Xiao *et al.*, 2017]. The details and statistics of these datasets are presented in the appendix. In addition, we compare our RDML method with eight state-of-the-art multi-view learning methods, including four evidence-based methods (i.e., TMC [Han *et al.*, 2020], ETMC [Han *et al.*, 2022b], ECML [Xu *et al.*, 2024a], and TMNR [Xu *et al.*, 2024b]), and four other uncertainty/confidence based methods (i.e., DUANets [Geng *et al.*, 2021], MMD [Han *et al.*, 2022a], QMF [Zhang *et al.*, 2023], and PDF [Cao *et al.*, 2024]. The details of all methods are shown in Appendix.

4.2 Implementation Details

Our experiments are conducted based on the PyTorch 2.4.1 framework with an Nvidia RTX 3090 GPU. For all datasets, 80% of the samples are used for training (for our method, these data are also used for pretraining), and 20% of the samples are used for testing. All experiments will be run 5 times,

Methods	Ref.	PIE	Scene	Leaves	NUS-WIDE	MSRC	Fashion
TMC	ICLR'21	91.85±0.23	67.71±0.30	86.81±2.20	35.67±1.37	92.38±2.78	95.40±0.40
ETMC	TPAMI'22	93.75±1.08	71.61±0.28	**98.44±0.40**	35.58±1.10	90.48±3.37	96.21±0.36
DUANets	AAAI'21	90.59±1.99	51.08±1.27	84.69±1.06	29.38±1.09	77.62±4.15	90.49±0.97
MMD	CVPR'22	94.41±2.01	65.72±1.38	69.05±1.18	28.21±5.11	98.10±2.33	9.58±0.61
QMF	ICML'23	88.82±1.82	65.24±1.80	95.19±1.53	42.33±2.56	94.76±3.81	98.81±0.16
ECML	AAAI'24	93.68±1.51	73.20±2.16	94.63±1.24	41.21±2.10	92.38±2.78	95.24±0.17
TMNR	IJCAI'24	89.71±1.61	66.24±2.05	89.31±1.80	36.75±1.71	90.00±1.78	94.31±0.45
PDF	ICML'24	90.88±1.36	69.61±1.72	98.00±0.51	43.83±1.73	93.33±3.50	98.85±0.14
RDML	Ours	**97.79±0.81**	**74.40±1.90**	97.94±1.09	**46.67±1.90**	**99.52±0.95**	**98.96±0.18**
△%		+3.38	+1.20	-0.50	+2.84	+1.42	+0.11

Table 1: Classification accuracy (%) on clean data, where the best and second best results are bolded and underlined, respectively. △% denotes the improvement of our method over the best baseline.

Methods	Ref.	PIE	Scene	Leaves	NUS-WIDE	MSRC	Fashion
TMC	ICLR'21	17.79±12.19	18.19±3.82	21.00±5.20	15.42±4.05	72.38±17.92	31.28±0.86
ETMC	TPAMI'22	40.29±19.39	13.51±3.36	73.44±17.91	16.71±7.95	83.81±5.30	74.94±0.41
DUANets	AAAI'21	0.59±0.29	1.07±0.36	0.38±0.61	1.58±0.50	2.86±1.78	1.74±1.13
MMD	CVPR'22	11.18±14.96	4.48±2.85	0.60±0.73	0.21±0.19	38.10±46.66	4.22±2.51
QMF	ICML'23	18.82±7.52	7.47±3.10	22.75±2.26	10.83±3.22	70.00±23.11	15.60±0.46
ECML	AAAI'24	41.62±9.92	6.67±3.02	54.40±11.10	16.83±8.12	80.00±19.08	18.28±2.99
TMNR	IJCAI'24	73.68±10.78	25.89±14.64	27.94±31.56	16.71±6.48	76.19±11.76	42.64±1.10
PDF	ICML'24	11.28±1.34	11.33±1.54	5.88±1.92	11.58±0.80	59.52±26.21	15.32±0.87
ETMC+AT	-	1.47+0.81	42.74+6.35	79.44+1.89	27.04+2.86	80.95+9.76	20.36+7.35
ECML+AT	-	7.50+11.07	33.00+4.87	78.81+7.68	25.38+4.36	60.48+29.30	90.70+1.33
TMNR+AT	-	71.32+11.44	38.08+16.64	49.62+3.39	20.33+2.28	73.81+3.01	72.16+0.69
PDF+AT	-	49.85+9.12	32.37+1.59	54.38+2.49	28.83+2.55	81.43+6.63	77.26+0.79
RDML	Ours	**88.97±4.08**	**48.67±3.97**	**87.25±5.90**	**29.83±3.99**	**89.05±7.62**	**91.19±0.69**
△%		+15.29	+5.93	+7.81	+1.00	+7.62	+0.49

Table 2: Classification accuracy (%) under adversarial attacks, where only a random view is attacked. AT denotes adversarial training.

and we will report the average performance and standard deviation based on the accuracy of each test (after the last training epoch). The pretraining epoch of $E_{pt}(\cdot)$ is 1000 with a batch size of 500, and the training epoch is 500 for the cleaning setting and 400 for the adversarial setting. The learning rate is selected from $[0.003, 0.005]$. $E^v_{pt}(\cdot)$ and $E^v_c(\cdot)$ are with the size of $[d_v, K]$. f^v_{EM} is with the size of $[K, d_v]$. And f^v_{EM} is with the size of $[d_v, d_v]$. Adam is used as the optimizer. The temperature μ of Gumbel softmax is set as 0.1. We use Projected Gradient Descent for random view attack. The number of attack iterations is 10 with a maximum perturbation range of $8/255$.

4.3 Experimental Results

To comprehensively evaluate the robustness of all methods, we conduct experiments on clean data and under adversarial attacks respectively. According to Table 1 and Table 2, we have following observations.

- For clean multi-view data, RDML achieves the best performance in most cases, with an average improvement of 1.41% compared to the best baselines. We attribute this to two reasons. First, the pretrained evidence extractor can provide good parameter initialization for the evidential classifier, avoiding falling into local optima prematurely during the training process. Second, evidential disentanglement learning, especially the combination of evidence-based disentanglement and View-specific evidential attention mechanism, is proficient in extracting multi-view features that are beneficial for classification and ignoring the interference of redundant information.
- RDML demonstrates significant advantages in AMVL. RDML has an average improvement of 6.36% compared to the second best methods (including AT based methods). Unlike existing evidence learning methods, RDML obtains a robust evidence extractor through perturbation-insensitive pretraining. Moreover, evidence-based disentanglement is adept at separating clean and adversarial features. Weak adversarial features are then repaired by the feature recalibration module, while the View-specific evidential attention can shield the interference of hard adversarial features on multi-view classification. Therefore, RDML is effective and robust under adversarial perturbations.
- Both evidence based methods and other types of methods are highly vulnerable to adversarial attacks. Although adversarial training can, to some extent, relieve the sensitivity of multi-view models to adversarial attacks in many cases, due to the lack of an appropriate

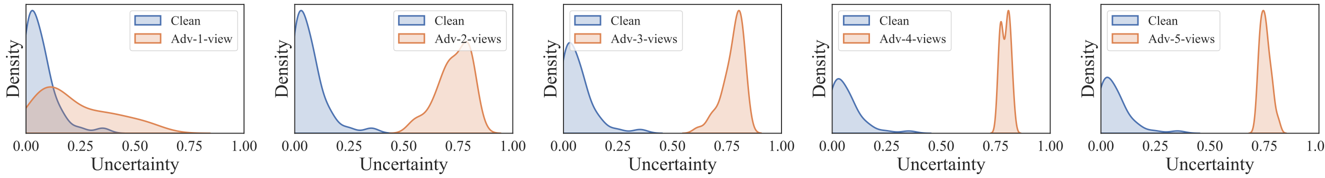


Figure 3: Density of estimated uncertainty on MSRC with different numbers of attacked views.

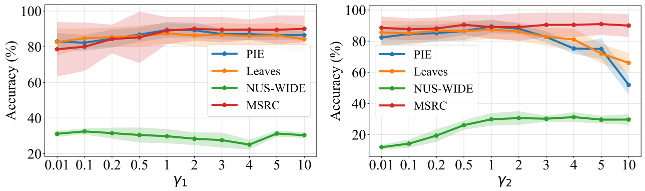


Figure 4: Classification accuracy (%) on four datasets with different γ_1 and γ_2 (one random view is attacked).

feature disentanglement mechanism, a large amount of adversarial and uninformative features are aggregated during multi-view fusion process, which impairs the classification performance.

4.4 Uncertainty Analysis

To study the effectiveness of the uncertainty estimation mechanism of our method under adversarial perturbation, we visualize the estimated uncertainty density in more adversarial scenarios. As we can observe from Fig.3, when only a few views are attacked, the estimated uncertainty can be well matched with that in the clean setting. However, when more views are attacked, the estimated uncertainty gradually increases. This phenomenon, on the one hand, indicates that the uncertainty estimation mechanism of our method is effective and robust under adversarial conditions. On the other hand, it also reveals that the uncertainty estimation mechanism is affected by the quality of the views. The worse the view quality, the more difficult it is to make accurate decisions. This further demonstrates the effectiveness of our evidential disentanglement learning and feature recalibration module, which are designed to decouple and repair adversarial and informative features, significantly improving the view quality.

4.5 Parameter Sensitivity Analysis

We conduct some experiments to study the influence of two key hyperparameters, γ_1 and γ_2, on model robustness. There are two points we can observe from Fig.4. (1) In most cases, the model performance rises slowly as the values of γ_1 and γ_2 increase. It reaches the optimal performance when the values are 1 or 2, and then declines slowly. Overall, our method is relatively insensitive to parameter changes, demonstrating its robustness. (2) When the values of γ_1 and γ_2 are too small (0.01) or too large (10), the model performance decreases significantly. On the one hand, this further validates the effectiveness of the evidential disentanglement loss and feature recalibration loss (γ_1 and γ_2 are their respective balancing coefficients). On the other hand, it also shows that overemphasizing either decoupling adversarial features or repairing adversarial information will undermine the model robustness. Striking an appropriate balance is the key to achieving better performance.

Ablation Module							Dataset		
$E_{pt}(\cdot)$	ED	FC	ATT	$\mathcal{L}_{ACL}$	$\mathcal{L}_{EDL}$	$\mathcal{L}_{FRL}$	PIE	Scene	Leaves
-	✓	✓	-	✓	✓	✓	51.62±14.46	48.34±5.14	67.56±5.03
✓	-	✓	✓	✓	-	✓	5.44±3.50	46.13±5.85	11.31±6.30
✓	✓	-	✓	✓	✓	-	82.79±2.35	41.58±6.91	84.25±6.92
✓	✓	✓	-	✓	✓	✓	87.79±4.12	48.18±3.72	86.06±5.88
✓	✓	✓	✓	-	✓	✓	86.47±3.80	47.54±4.55	86.81±4.92
✓	✓	✓	✓	✓	-	✓	83.68±4.14	48.16±4.30	83.13±6.34
✓	✓	✓	✓	✓	✓	-	81.76±6.21	46.98±4.70	86.06±7.58
✓	✓	✓	✓	✓	✓	✓	**88.97±4.08**	**48.67±3.97**	**87.25±5.90**

Table 3: Ablation experiments (classification accuracy (%)) on PIE, Scene, and Leaves with one randomly attacked view. ED, FC, and ATT denote evidential disentanglement learning, feature recalibration, and evidential attention mechanism, respectively.

4.6 Ablation Study

We conduct ablation experiments to verify the effectiveness of key components of RDML. As shown in Table 3, after removing the pretrained $E_{pt}(\cdot)$ and evidence-based disentanglement respectively (the evidence attention mechanism and $\mathcal{L}_{EDL}$ then get invalid automatically), the model robustness significantly decreases. This indicates that: (1) perturbation-insensitive pretraining is highly effective in helping the evidential neural network resist adversarial attacks. (2) evidence-based decoupling can effectively strip the adversarial perturbations from view representations, and effectively reduce the interference of adversarial and uninformative features. After removing the feature recalibration and evidential attention respectively, the model performance shows a decline to varying degrees in most cases. In addition, we also explore the effectiveness of three proposed losses. The results prove that the three losses, especially $\mathcal{L}_{EDL}$ and $\mathcal{L}_{FRL}$, are able to improve the model robustness in most cases.

5 Conclusion

In this paper, we reveal and study the adversarial unreliability problem in trustworthy multi-view learning. To this end, we propose a novel multi-view learning framework, namely Reliable Disentanglement Multi-view Learning. Specifically, our RDML designs an evidential disentanglement learning to separate clean and adversarial features, and this process is guided by a pretrained evidence extractor. To mitigate the interference of adversarial features on multi-view decision, an adversarial feature recalibration module and an evidential attention mechanism are proposed. Various experiments conducted on six datasets show the effectiveness and robustness of RDML against view adversarial attacks.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant Nos. 62133012, 61936006, 62103314, 62073255, 62303366), and the Sichuan Science and Technology Miaozi Program (Grant no. MZGC20240057).

References

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

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009.

[Cope *et al.*, 2013] James Cope, Thibaut Beghin, Paolo Remagnino, and Sarah Barman. One-hundred plant species leaves data set. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C5RG76.

[Fei-Fei and Perona, 2005] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 2, pages 524–531. IEEE, 2005.

[Geng *et al.*, 2021] Yu Geng, Zongbo Han, Changqing Zhang, and Qinghua Hu. Uncertainty-aware multi-view representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7545–7553, 2021.

[Goodfellow *et al.*, 2015] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[Gross *et al.*, 2010] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010.

[Han *et al.*, 2020] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2020.

[Han *et al.*, 2022a] Zongbo Han, Fan Yang, Junzhou Huang, Changqing Zhang, and Jianhua Yao. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20707–20717, 2022.

[Han *et al.*, 2022b] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):2551–2566, 2022.

[Ilyas *et al.*, 2019] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

[Jang *et al.*, 2022] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2022.

[Jsang, 2018] Audun Jsang. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated, 2018.

[Kim *et al.*, 2023] Woo Jae Kim, Yoonki Cho, Junsik Jung, and Sung-Eui Yoon. Feature separation and recalibration for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8183–8192, 2023.

[Kurakin *et al.*, 2018] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[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.

[Liu *et al.*, 2024] Jun Liu, Jiantao Zhou, Jiandian Zeng, and Jinyu Tian. Difattack: Query-efficient black-box adversarial attack via disentangled feature space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3666–3674, 2024.

[Long *et al.*, 2022] Teng Long, Qi Gao, Lili Xu, and Zhangbing Zhou. A survey on adversarial attacks in computer vision: Taxonomy, visualization and future directions. *Computers & Security*, 121:102847, 2022.

[Madry *et al.*, 2018] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[Peng *et al.*, 2019] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. Comic: Multi-view clustering without parameter selection. In *International conference on machine learning*, pages 5092–5101. PMLR, 2019.

[Qin *et al.*, 2024] Yalan Qin, Xinpeng Zhang, Shui Yu, and Guorui Feng. A survey on representation learning for multi-view data. *Neural Networks*, page 106842, 2024.

[Sensoy *et al.*, 2018] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[Sun *et al.*, 2024] Yuan Sun, Yang Qin, Yongxiang Li, Dezhong Peng, Xi Peng, and Peng Hu. Robust multi-view clustering with noisy correspondence. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[Wang *et al.*, 2022] Jia Wang, Chengyu Wang, Qiuzhen Lin, Chengwen Luo, Chao Wu, and Jianqiang Li. Adversarial attacks and defenses in deep learning for image recognition: A survey. *Neurocomputing*, 514:162–181, 2022.

[Xiao *et al.*, 2017] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[Xu *et al.*, 2016] Jinglin Xu, Junwei Han, and Feiping Nie. Discriminatively embedded k-means for multi-view clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5356–5364, 2016.

[Xu *et al.*, 2024a] 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.

[Xu *et al.*, 2024b] Cai Xu, Yilin Zhang, Ziyu Guan, and Wei Zhao. Trusted multi-view learning with label noise. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 5263–5271. ijcai.org, 2024.

[Xu *et al.*, 2024c] Jie Xu, Yazhou Ren, Xiaolong Wang, Lei Feng, Zheng Zhang, Gang Niu, and Xiaofeng Zhu. Investigating and mitigating the side effects of noisy views for self-supervised clustering algorithms in practical multi-view scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22957–22966, 2024.

[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.*, 2020] Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. Deep partial multi-view learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2402–2415, 2020.

[Zhang *et al.*, 2023] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pages 41753–41769. PMLR, 2023.

[Zhang *et al.*, 2024] Yufeng Zhang, Jianxing Yu, Yanghui Rao, Libin Zheng, Qinliang Su, Huaijie Zhu, and Jian Yin. Domain adaptation for subjective induction questions answering on products by adversarial disentangled learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 9074–9089, 2024.

Appendix

This supplementary material provides a comprehensive understanding of our RDML method. Specifically, we mainly introduce the algorithm procedure, the details of used datasets and compared methods, and more experimental analysis to support our research.

A Algorithm Procedure

To better show the details of our proposed method, we give the workflow of our RDML in Algorithm 1.

Algorithm 1 The workflow of RDML

Input: Multi-view data x; pretraining epoch e_{pt} and training epoch e_t; evidence extractor $E_{pt}(\cdot)$ and $E_c(\cdot)$; evidence mapping layer $f_{EM}(\cdot)$ and evidential classifier $f_{EC}(\cdot)$
Output: Joint opinion.

```
..............................Pretraining stage..............................
Randomly initialize the parameters of $E_{pt}(\cdot)$.
while not reach the last epoch $e_{pt}$ do
    Construct the adversarial instance with random view attacks by Eq. (1);
    Train the evidence extractor $E_{pt}(\cdot)$ with mixed data for evidence-based multi-view classification using Eq. (2).
end while
Share the parameters of $E_{pt}(\cdot)$ with $E_c(\cdot)$, and freeze the parameters of $E_{pt}(\cdot)$.
................................Training stage................................
Randomly initialize the parameters of $f_{EM}(\cdot)$ and $f_{EC}(\cdot)$.
while not reach the last epoch $e_t$ do
    Construct the adversarial instance with random view attacks by Eq. (1);
    Decouple view features with the guidance of $E_{pt}(\cdot)$ via Eq. (9)-Eq. (12);
    Recalibrate view features using Eq. (14);
    Construct view-level evidential attention to further mitigate the interference of adversarial perturbations via Eq. (16) and Eq. (17);
    Construct joint opinion using Dempster combination rule via Eq. (18);
end while
```

B Experimental details

B.1 Datasets

Dataset	Class	Size	Dimensionality
PIE	68	680	484/256/279
Scene	15	4485	20/59/40
Leaves	100	1600	64/64/64
NUS-WIDE	12	2400	64/144/73/128/225
MSRC	7	210	24/576/512/256/254
Fashion	10	10000	784/784/784

Table 4: Details of datasets

We evaluate the performance of our propose method on the following multi-view datasets. **PIE** encompasses 680 facial images sourced from 68 subjects. Three kinds of views (intensity, LBP and Gabor) are selected in our experments. **Scene** dataset comprises 4485 images, which are categorized into 15 distinct indoor and outdoor scene classes. Three types of features are selected: GIST, PHOG, and LBP. **Leaves** dataset is constituted by 1600 leaf samples, which are collected from 100 diverse plant species. We extract three types of features as views: shape descriptors, fine-scale edges, and texture histograms. **NUS-WIDE** dataset comprises 269648 images from 81 concepts. We select 200 images from each of the top 12 classes, with a total of five types of views: CH, CM, CORR, EDH and WT. **MSRC-v5** from Microsoft Research in Cambridge contains 210 images and 7 classes with coarse pixel-wise labeled images. Five types of views are extracted: CM, HOG, GIST, LBP and CENT. **Fashion** comprises of grayscale images of 70,000 fashion products from 10 categories. We sample 1000 images from each class, with a total of three types of views. Details of these datasets are shown in Table 4.

B.2 Baselines

To verify the effectiveness and robustness of our method, we compare RDML with eight state-of-the-art multi-view methods. **TMC** applies evidence theory to multi-view learning, dynamically fusing various views at the evidence level. **ETMC** is an improved version of TMC. It enhances the performance of TMC by adding a new view which is the concatenation of all original views. **DUA-Nets** capitalizes on reversal networks to amalgamate the intrinsic information sourced from diverse views, ultimately transforming them into a unified representation. **MMD** dynamically evaluates feature and modality informativeness with specific strategies and induces a transparent fusion algorithm. **QMF** utilizes uncertainty-aware weighting and a sampling-based regularization technology to enhance correlation, aiming to achieve reliable and robust multimodal fusion. **ECML** proposes a new multi-view opinion fusion method and a conflict measurement method to solve the problem of aggregating conflicting opinions. **TMNR** proposes a noise correlation matrix, through which the Dirichlet parameters are updated to mitigate the interference of label noise. **PDF** derives the predictable Collaborative Belief with Mono- and Holo-Confidence to reduce the generalization error upper bound and further proposes a relative calibration strategy.

C Multi-view Attack Analysis

In order to further verify the robustness of our method, we conducted experiments in more difficult scenarios. As shown in Fig.5, as the number of attacked views increases, the performance of all methods is inevitably impaired, while our method can always maintain the best adversarial robustness. We attribute this to the combination of the adversarial-insensitive pretrained evidence extractor and the evidence-based disentanglement mechanism. The former can ensure the acquisition of effective evidences in a strong adversarial environment, and the latter can strip away adversarial pertur-

Method	PIE			MSRC		
	$eps = 8/255$	$eps = 0.05$	$eps = 0.1$	$eps = 8/255$	$eps = 0.05$	$eps = 0.1$
RDML	88.97±4.08	88.24±4.88	94.26±1.18	89.05±7.62	88.10±7.38	89.52±5.55
TMNR	73.68±10.78	68.97±15.92	59.12±22.26	76.19±11.76	70.48±18.36	73.33±15.36
PDF	11.28±1.34	8.53±4.38	8.53±4.78	59.52±26.21	51.90±32.38	51.43±35.04

Table 5: eps is the maximum perturbation range. Larger values indicate stronger attacks. $8/255$ is the empirically default setting. One random view is attacked.

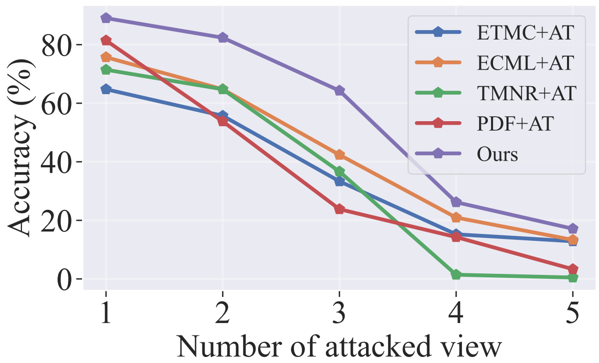


Figure 5: Classification accuracy (%) on MSRC with different numbers of attacked views.

bations and reduce their interference with the distribution of clean features.

D Attack Intensity

As shown in Table 5, TNMR and PDF exhibit clear performance degradation as the attack strength increases. Benefiting from perturbation-insensitive pretraining, RDML not only maintains robustness but even achieves superior performance under stronger attacks, further demonstrating its effectiveness.

E Parameter Freezing

Since neural networks are highly sensitive to perturbations, we freeze the parameters of the evidence extractor after pretraining to ensure model stability and robustness. As shown in Table 6, parameter freezing significantly enhances model robustness.

Method	PIE	Leaves	MSRC
RDML	88.97±4.08	87.25±5.90	89.05±7.62
w/o freezing	78.24±7.03	84.56±6.33	86.67±8.73

Table 6: Model performance with and without parameter freezing. One random view is attacked.

F Gumbel Softmax

In evidential disentanglement learning, we transform evidential representations into feature-level masks. To ensure differentiability, we introduce Gumbel softmax. So each value of the mask is non-negative and less than 1. The motivation of introducing the Gumbel distribution is to provide a differentiable reparameterization method for sampling discrete random variables, thereby addressing the issue that discrete variables cannot directly propagate gradients during backpropagation.