Enhancing Multi-View Classification Reliability with Adaptive Rejection

Wei Liu[1], **Yufei Chen**[1 *], **Xiaodong Yue**[2]

[1] School of Computer Science and Technology, Tongji University, Shanghai, China
[2] Artificial Intelligence Institute of Shanghai University, Shanghai University, Shanghai, China
ldachuan@outlook.com, yufeichen@tongji.edu.cn, yswantfly@shu.edu.cn

Abstract

Multi-view classification based on evidence theory aims to enhance result reliability by effectively quantifying prediction uncertainty at the evidence level, particularly when dealing with low-quality views. However, these methods face limitations in real-world applications due to the sensitivity of estimated uncertainty to view distribution, leading to two main issues: 1) difficulty in making clear judgments about whether to trust predictions based on vague uncertainty scores, and 2) the potential negative impact of integrating information from low-quality views on multi-view classification performance. Both limitations compromise the reliability of multi-view decisions. To address these challenges, we introduce an adaptive rejection mechanism based on estimated uncertainty, which is free of data distribution constraints. By integrating this adaptive rejection mechanism into the fusion of multiple views, our method not only indicates whether predictions should be adopted or rejected at the view level but also enhances classification performance by minimizing the impact of unreliable information. The effectiveness of our method is demonstrated through comprehensive theoretical analysis and empirical experiments on various multi-view datasets, establishing its superiority in enhancing the reliability of multi-view classification.

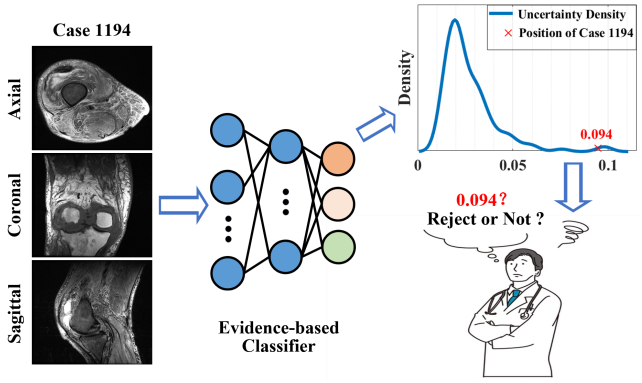

Figure 1: The evidence-based classifier is trained on the MR-Net dataset, which includes real-world knee MRI scans captured from three views: Axial, Coronal, and Sagittal. Case 1194 is a test sample with noise present in all three views. The blue line indicates the density of estimated uncertainty for testing samples. The red "×" denotes the position of Case 1194 within the uncertainty density distribution of the test samples, showing a small uncertainty value of 0.094.

Introduction

Uncertainty-aware Multi-view Classification (UMVC) has recently become crucial in multi-view deep learning by assessing uncertainty levels in predictions, thus informing decision-makers about when to trust these predictions (Zou et al. 2023; Zhou et al. 2023). Such uncertainties typically arise from low-quality data, including conflicts among multiple views or the presence of outliers (Zhang et al. 2024b). Within UMVC, the framework based on evidence theory has emerged as a promising approach (Han et al. 2021; Liu et al. 2022; Han et al. 2023; Liu et al. 2023; Xu et al. 2024; Lu et al. 2024). This framework leverages distinct views at the evidence level, providing a reliable means for decision fusion. It has been significantly applied in various safety-critical fields, such as medical diagnosis (Wang et al. 2023) and autonomous driving (Patrikar et al. 2024).

Despite their clear benefits, these approaches face limitations that may compromise the reliability of multi-view deep learning. Specifically, existing methods typically provide an uncertainty degree to indicate whether a prediction can be adopted. However, the magnitude of this uncertainty is sensitive to different multi-view data distributions and even varies across different view distributions within a single dataset, resulting in inconsistent levels of uncertainty. This variability poses a critical challenge: how to effectively distinguish between predictions from normal data and those from low-quality data across various data distributions. Consequently, two main issues arise: **1)** the integration of unreliable evidence from low-quality views can degrade the performance of multi-view classification, and **2)** The principle in informatics, as stated by Shannon, that "the essence of information is to eliminate uncertainty", cited in (Burgin 2002), is followed by most evidence-based multi-view methods, typically resulting in reduced uncertainty during multi-view fusion. This low uncertainty diminishes the distinction between the uncertainty scores of reliable and un-

*Corresponding author

reliable predictions, complicating the decision-making process regarding whether a prediction should be trusted or rejected. Such difficulties can potentially increase risks in real-world applications, particularly in high-stakes scenarios (e.g., medical diagnosis) that require critical decision-making.

For example, as illustrated in Figure 1, we estimate prediction uncertainty using the evidence-based method (Liu et al. 2022) across the MRNet dataset (Bien et al. 2018), which comprises real-world knee Magnetic Resonance Imaging (MRI) scans captured from three views. Case 1194 serves as a test sample characterized by significant noise and a lack of useful information. Despite the estimated uncertainty density for Case 1194 is located in the tail, its specific value of 0.094, which is relatively low, may lead to ambiguity for doctors regarding the reliability of the prediction. In practice, the desired reliable output for physicians is not only to ensure accurate diagnoses for normal multi-view data but also to provide a clear judgment on the rejection of diagnoses for low-quality samples.

To achieve this goal, an intuitive approach is using the classification with *Reject Option*, also called *Selective Classification* (SC), a method that has been around for 70 years (De Stefano, Sansone, and Vento 2000; Geifman and El-Yaniv 2017). This method mainly optimizes the selective risk through a threshold to determine whether the classification result should be rejected. However, there are several challenges with existing SC methods: **(1)** Existing SC methods typically either depend on a fixed threshold, such as the average threshold of a classifier ensemble (Varshney 2011), which cannot adaptively apply to complex data distributions, or optimize the threshold through reject risk optimization (Franc, Prusa, and Voracek 2023), leading to varying performance outcomes due to differences in model structure. **(2)** They hardly provide a theoretical guarantee that the risk of rejection for the prediction of low-quality samples is bounded by a small marginal probability. **(3)** There is limited SC work addressing the handling of multi-view data, which should take into account the rejection relation among different views.

The above gaps motivate us to devise a novel method called Reliable Multi-view Classification with Adaptive Rejection (RAR). Specifically, we first aim to construct a distribution-free rejection mechanism with p-values to provide an adaptive rejection option at the view level, indicating whether we should reject the evidence collected from a specific view. Then, we introduce the adaptive rejection option into the multi-view fusion module, which can eliminate the effect of unconfident results and improve classification performance. During inference, we can make a clear judgment on when to trust the multi-view result in terms of the rejection option associated with the overall uncertainty score, thereby enhancing reliability in multi-view deep learning. In summary, the contributions of this paper are as follows: (1) We introduce a novel multi-view model with adaptive rejection aimed at enhancing the reliability of multi-view decisions, providing a clear judgment on whether the decision should be kept or rejected; (2) The proposed rejection mechanism is distribution-free and can be adaptively applied to various multi-view data structures. The devised multi-view fusion strategy with rejection can eliminate the effect of unconfident results and improve classification performance; (3) Theoretical and empirical analyses demonstrate the effectiveness of the proposed model in terms of accuracy, and reliability.

Related Works

Evidence-based Multi-view Classification: Up till now, numerous multi-view learning approaches using deep neural networks (DNNs) have emerged (Sun, Dong, and Liu 2020; Liang et al. 2021, 2024; Zhang et al. 2024a; Jiang et al. 2024; Zhang et al. 2022; Liu, Chen, and Yue 2024), which aims to learn a shared representation of multiple information from different types of views with DNNs. While multi-view deep learning demonstrates notable effectiveness, the performance is affected by the presence of low-quality multi-view data, which introduce risks in real-world scenarios. In response to this challenge, recently, several uncertainty-aware multi-view classification methods, particularly evidence-based approaches (Han et al. 2021; Liu et al. 2022; Han et al. 2023; Liu et al. 2023; Xu et al. 2024; Lu et al. 2024; Xu et al. 2023; Liu et al. 2024), have been proposed. These methods use subjective logic to model view-specific predictions as multinomial opinions and then combine them using appropriate belief fusion rules based on the context (Jøsang 2018). Specifically, an extension of the Dempster-Shafer combination rule (Shafer 1976) was first adopted by (Han et al. 2021, 2023) in multi-view classification. Other fusion rules, such as Cumulative Belief Fusion (Liu et al. 2022, 2023) and Averaging Belief Fusion (Xu et al. 2024), have also been explored. Despite their success in UMVC, these methods fall short in enhancing reliability in real-world environments. The key issue is that the magnitude of uncertainty is often unmeasurable due to the variability of multi-view distributions, leading to the following problems: **1)** Poor classification performance with vague uncertainty degrees. **2)** An inability to provide clear judgments on whether the prediction is reliable. In contrast, our method introduces an adaptive rejection mechanism into the fusion of multiple views, improving the performance of multi-view results and providing clear judgments on whether the prediction should be rejected, thereby boosting reliability in multi-view classification.

Classification with Reject Option: Classification with a reject option, also known as selective classification, aims to abstain from making predictions on examples that the model is likely to misclassify (De Stefano, Sansone, and Vento 2000; Geifman and El-Yaniv 2017). A straightforward and popular technique is to use some confidence measure to select the most certain examples (Zhang et al. 2023). If the underlying confidence degree of the predicted class is below a certain threshold, the prediction will abstain. However, the threshold is either fixed (Varshney 2011) or depends on selective risk optimization (Franc, Prusa, and Voracek 2023), which is inflexible for variations in model structure and data distribution, leading to poor performance. Moreover, there has been limited focus on multi-view classification with a

reject option, especially one that incorporates an adaptive threshold applicable to various data distributions within a low rejection risk. This gap motivates the development of the work presented in this manuscript.

Method

This section outlines the notations and problem statement, followed by a detailed description of the proposed method and its key components.

Problem Statement

Consider a multiclass classification task on data $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where $x = \{x^v\}_{v=1}^V$ represents data from V different views, and $y \in \mathcal{Y}$ represents the corresponding labels. The multi-view training data is denoted as $\mathcal{D}_{train} = \{x_i, y_i\}_{i=1}^n$. UMVC learns a multi-view model f on $\mathcal{D}_{train}$, mapping x to a top-1 prediction $\hat{y}$ with an associated uncertainty degree u. However, *the magnitude of uncertainty is sensitive to the distributions of multi-view data*, which can adversely affect classification performance and complicate the assessment of prediction reliability based on ambiguous degrees of uncertainty. To address this, we introduce an adaptive rejection mechanism into the multi-view fusion process.

Preliminaries on Evidence Theory

Given the promising application of evidence-based frameworks in UMVC (Han et al. 2021; Liu et al. 2022; Han et al. 2023; Liu et al. 2023; Xu et al. 2024), which utilize evidence theory (Shafer 1976) and subjective logic (Jøsang 2018) to handle K-class classification by assigning belief masses to individual class labels and computing epistemic uncertainty, we adopt this approach as our foundational framework.

Formally, let $f_v(x^v; \theta)$ denote the output of the penultimate (logits) layer of the view-specific neural network for the sample x^v in a single view. Rather than using the standard *softmax* function to predict a single categorical estimation, we employ an alternative activation function $a(\cdot)$ (e.g., ReLU) to capture non-negative evidence $\boldsymbol{e}^v = a(f_v(x^v; \theta))$, where $\boldsymbol{e}^v = [e_1^v, \ldots, e_K^v]$ over K categories. Then, we define:

$$b_k^v = e_k^v / S^v, u^v = 1 - \sum_{k=1}^{K} b_k^v = K / S^v, \tag{1}$$

where $S^v = \sum_{k=1}^{K} e_k^v + K$. Here, b_k^v represents the belief mass of the k^{th} class based on the collected evidence, and u^v indicates the prediction uncertainty, which reflects the lack of total evidence. However, managing the magnitude of uncertainty across different data distributions poses challenges, necessitating the introduction of an adaptive rejection mechanism.

Adaptive Rejection Mechanism

To make a clear rejection decision, an intuitive method is to learn a rejection function $g : \mathcal{X} \to \{0, 1\}$ using an uncertainty threshold u_λ:

$$g(x) = \begin{cases} 1 & u(x) < u_\lambda, \\ 0 & else, \end{cases} \tag{2}$$

where $u(x)$ is the estimated uncertainty of any test point x. A prediction is made if $g(x) = 1$, and otherwise, it is rejected. However, two main challenges arise: **1)** determining an adaptive threshold applicable across various data distributions, and **2)** controlling the risk of false rejections. To address these challenges, we treat the rejection decision as a multiple-testing problem. Given a test dataset $\mathcal{D}_{test} = \{x_j, y_j\}_{j=1}^m$ with m samples, rejection is viewed as testing the following *random* hypotheses:

$$H_j : u(x_j) \leq u_\lambda, j = 1, \ldots, m. \tag{3}$$

We define $H_0 = \{j : H_j \; is \; true\}$ as the set of *null* hypotheses. Rejecting the *null* hypothesis indicates that $u(x_j)$ exceeds the threshold u_λ, leading to the rejection of the prediction for test sample x_j.

Given the absence of a true label indicating whether a prediction is unreliable in practice, we aim to control the risk of rejection using the statistical error rate (Benjamini and Hochberg 1995; Angelopoulos et al. 2021). This can be formally achieved by computing p-values p_j for any *null* hypothesis H_j based on an independent calibration dataset $\mathcal{D}_{cal}$ (with the same distribution as $\mathcal{D}_{train}$):

$$p_j = \frac{1 + card\{x_s \in \mathcal{D}_{cal} : u(x_s) \leq u(x_j)\}}{|\mathcal{D}_{cal}| + 1}, \tag{4}$$

where p_j indicates the extent to which the instance is typical of reliable results. A smaller p-value suggests a higher possibility of false rejections. Moreover, p-values satisfy the following property (Nouretdinov et al. 2001):

$$\mathbb{P}(p_j \leq \tau) \leq \tau, \tag{5}$$

with $\tau \in (0, 1)$. When τ is considered as the chosen risk tolerance, this property ensures that if we control the possibility of making false rejections does not exceed τ for any test samples, we will control the rejection risk at the τ-level.

Thus, the problem reduces to finding valid p-values that balance performance with Eq. (5). To achieve this, we define the following:

Definition 1. *For $x_s \in \mathcal{D}_{cal}$, $s = 1, \ldots, l.$, let u_s denote $u(x_j)$. The random uncertainty scores $u_{(1)}, u_{(2)}, \cdots, u_{(l)}$ are exchangeable with order statistics $u_{(1)} \leq u_{(2)} \leq \cdots \leq u_{(l)}$. For any fixed $\tau \in (0, 1)$, u_λ can be defined as $u_\lambda = u_{(\lceil (1-\tau)(1+l) \rceil)}$. Note that when l is too small, specifically when $\tau < 1/(l+1)$, u_λ will approach infinity.*

We then establish the following proposition:

Proposition 1. *For any size of $\mathcal{D}_{cal}$, the following inequality holds for any null hypothesis with $\tau \in (0, 1)$:*

$$\mathbb{P}(\mathbb{P}(p_j \leq \tau) \leq \tau) \geq 1 - \delta,$$

for any $\delta \in (0, 1)$, where δ is the significance level.

Definition 1 provides an adaptive threshold u_λ, which aids in the development of a rejection function $g(x)$. This function is independent of data distribution and controls the risk of rejection based on a user-specified τ (e.g., $\tau = 0.001$).

Reliable Fusion with Adaptive Rejection

To generate the final prediction, we model the view-specific predictions as multinomial opinions, denoted as $\mathcal{O}^v = \left\{\boldsymbol{b}^v = \{b_k^v\}_{k=1}^K, u^v\right\}$, and then combine them using appropriate dynamic fusion rules. In contrast to the existing evidence-based multi-view methods, we introduce a rejection function into the fusion of multiple views and propose the following fusion strategy.

Definition 2. ***(Multi-view fusion strategy).*** *Given an opinion $\mathcal{O}_i^v$ for an arbitrary x_i^v, let $\tilde{e}_{ik}^v = e_{ik}^v \cdot g\left(x_i^v\right)$ be the transformation of collected evidence e_{ik}^v in the rejection mechanism. We define $\tilde{b}_{ik}^v = b_{ik}^v \circ g\left(x_i^v\right) = \frac{e_{ik}^v + \tilde{e}_{ik}^v}{\sum_{k=1}^K \left(e_{ik}^v + \tilde{e}_{ik}^v\right) + 2K}$, and $\tilde{u}_i^v = 1 - \sum_{k=1}^K \tilde{b}_{ik}^v$ using Eq. (1), where $\circ$ denotes a composition. Let $\mathcal{O}_i^1$ and $\mathcal{O}_i^2$ be opinions of any two views. The fusion $\mathcal{O}_i^1 \oplus \mathcal{O}_i^2$ can be achieved by the following rule:*

$$\begin{aligned} b_{ik}^1 \oplus b_{ik}^2 &= \frac{\tilde{b}_{ik}^1 \tilde{u}_i^2 + \tilde{b}_{ik}^2 \tilde{u}_i^1}{\tilde{u}_i^1 + \tilde{u}_i^2 - \tilde{u}_i^1 \tilde{u}_i^2}. \\ u_i^1 \oplus u_i^2 &= \frac{\tilde{u}_i^1 \tilde{u}_i^2}{\tilde{u}_i^1 + \tilde{u}_i^2 - \tilde{u}_i^1 \tilde{u}_i^2}. \end{aligned} \tag{6}$$

Without loss of generality, let $\mathcal{O}_i = \left\{\boldsymbol{b}_i = \{b_{ik}\}_{k=1}^K, u_i\right\}$ be the opinion fused from V views, we can easily expand the above rule to multiple views fusion:

$$\mathcal{O}_i = \oplus_{v=1}^V \mathcal{O}_i^v = \mathcal{O}_i^1 \oplus \cdots \oplus \mathcal{O}_i^V. \tag{7}$$

Following Definition 2, we can also derive the multi-view evidence $\boldsymbol{e}_i = [e_{i1}, \ldots, e_{iK}]$ over K categories using Eq.(1). Notably, the proposed multi-view fusion strategy has the following properties:

Proposition 2. *If $u_i^v > u_\lambda^v$ for $\forall v \in [1, \ldots, V]$, the uncertainty of the integrated opinion $\mathcal{O}_i$ satisfies $u_i > u_\lambda$. If $\exists v \in [1, \ldots, V]$, $u_i^v \leq u_\lambda$, then $u_i \leq u_\lambda$.*

Proposition 2 ensures a more intuitive outcome. It guarantees that when the evidence provided by each view is of high uncertainty and should be rejected, the integrated evidence from these views will also be rejected. Conversely, if any view presents reliable evidence that should not be rejected, the integrated evidence from these views will be incorporated into the final decision-making process.

Proposition 3. *Let $\mathcal{O}_i^o$ be an original opinion with a correct prediction, and let $\mathcal{O}_i^v$ be an arbitrary opinion with $u_i^v > u_\lambda^v$. The evidence from $\mathcal{O}_i^o$ will be given more weight when integrating $\mathcal{O}_i^v$ into $\mathcal{O}_i^o$. Moreover, suppose t is the index of the ground-truth class, $b_{it} \geq b_{ik}$ always holds under the condition $b_{it}^v = b_{iz}^v$, where z is the largest belief mass in $\{b_{ik}^v\}_{k=1}^K$.*

Proposition 3 guarantees that fusion with adaptive rejection can mitigate the impact of unreliable evidence, leading to better multi-view fusion classification performance. Together, these propositions provide a theoretical guarantee of reliability in multi-view classification.

Learning to Form Opinions

In the proposed training algorithm, we follow (Şensoy, Kaplan, and Kandemir 2018), using the following loss to optimize the network parameters:

$$\mathcal{L}_{overall}\left(\theta\right) = \sum_{i=1}^n \left(\ell_{acc}\left(\theta\right)_i + \ell_{kl}\left(\theta\right)_i\right), \tag{8}$$

where $\ell_{acc}\left(\theta\right)_i$ represents the classification loss term, and $\ell_{kl}\left(\theta\right)_i$ is the KL-divergence regularization term, used to penalize those incorrect evidence by shrinking it to 0.

Specifically, for a training sample x_i with a one-hot label $\boldsymbol{y}_i$, we treat the Dirichlet distribution $Dir\left(\boldsymbol{p}_i \left| \boldsymbol{\alpha}_i \right.\right)$ as a prior on the likelihood, where $\boldsymbol{\alpha}_i = \boldsymbol{e}_i + 1$, and obtain the negative log likelihood loss:

$$\begin{aligned} \ell_{acc}\left(\theta\right)_i &= \mathbb{E}_{\boldsymbol{p}_i \sim Dir\left(\boldsymbol{p}_i \left| \boldsymbol{\alpha}_i \right.\right)} \left[-\sum_{k=1}^K \log\left(p_{ik}^{y_{ik}}\right) \right] \\ &= \sum_{k=1}^K y_{ik} \left(log\left(S_i\right) - log\left(e_{ik} + 1\right) \right), \end{aligned} \tag{9}$$

where $S_i = \sum_{k=1}^K e_{ik} + K$, and $\boldsymbol{p}_i = [p_{i1}, \ldots, p_{iK}]$ is the expectation of the $Dir\left(\boldsymbol{p}_i \left| \boldsymbol{\alpha}_i \right.\right)$. Then we have:

$$\ell_{kl}\left(\theta_i\right) = \lambda_t \cdot KL\left[Dir\left(\boldsymbol{p}_i \left| \tilde{\boldsymbol{\alpha}}_i \right.\right) || Dir\left(\boldsymbol{p}_i \left| \mathbf{1} \right.\right)\right], \tag{10}$$

where $\tilde{\boldsymbol{\alpha}}_i = \boldsymbol{y}_i + \left(1 - \boldsymbol{y}_i\right) \odot \boldsymbol{\alpha}_i$, with $\odot$ denoting element-wise product. The term $Dir\left(\boldsymbol{P}_i \left| \mathbf{1} \right.\right)$ represents the uniform Dirichlet distribution, and λ_t is an annealing factor. We define $\lambda_t = \lambda_0 \exp\left\{-\left(\ln \lambda_0 / T\right) t\right\}$, where λ_0 is a small positive constant, t is the current training epoch, and T is the total number of training epochs. As t approaches T, λ_t will increase exponentially from λ_0 to 1.

Computational Complexity Analysis

Our method presents a generalized framework applicable to various DNNs. Let the training complexity of any multi-view deep neural network be $O\left(T_{train}\right)$, and the complexity of one inference pass be $O\left(T_{inf}\right)$. The final computational complexity is given by:

$$O\left(T_{train}\right) + O\left(l \log\left(l\right)\right) + O\left(m \cdot T_{inf}\right), \tag{11}$$

where l and m are the sizes of the calibration and inference datasets, respectively, and $O\left(l \log\left(l\right)\right)$ indicates the complexity of threshold calculation. In most practical scenarios, T_{train} will dominate, particularly for large-scale DNNs, unless the dataset is exceptionally large or the DNNs are very shallow or simple.

Experiments

Experimental Setup

Datasets: We conducted experiments on six real-world multi-view datasets as follows: **ANIMAL** (Lampert, Nickisch, and Harmeling 2013) includes 10,158 images from 50 classes, with deep features extracted using DECAF and VGG19 as two views. **HAND** (van Breukelen et al. 1998)

Components			ANIMAL	HAND	CUB	SCENE	MRNet	CAL
Fusion	Rejection	ℓ_{kl}	ACC (%)↑	ACC (%)↑	ACC (%)↑	ACC (%)↑	ACC (%)↑	ACC (%)↑
✗	✗	✗	90.46±0.60	94.44±0.83	87.28±0.46	69.08±1.05	87.87±0.31	95.54±0.99
✓	✗	✗	94.17±0.20	98.23±0.35	93.74±1.31	73.95±1.15	92.47±0.39	96.93±0.79
✓	✗	✓	94.91±0.33	98.78±0.28	94.37±0.17	74.03±1.60	93.00±0.31	97.03±1.01
✓	✓	✗	95.32±0.59	98.98±0.39	95.01±1.01	74.70±1.11	94.17±0.11	97.62±0.58
✓	✓	✓	**95.65±0.36**	**99.11±0.27**	**95.19±1.59**	**75.69±1.35**	**95.00±0.10**	**97.62±0.37**

Table 1: Classification performance with the corresponding component.

Method	ANIMAL	HAND	CUB	SCENE	MRNet	CAL
DCCA (Andrew et al. 2013)	81.10±1.22	94.01±1.20	78.12±1.00	51.21±1.67	87.21±2.79	80.00±0.77
DCCAE (Wang et al. 2015)	84.97±0.21	97.01±0.23	80.42±0.88	52.12±0.31	88.23±0.11	88.11±0.79
CPM-Nets (Zhang et al. 2020)	85.21±0.20	94.00±1.10	85.67±0.02	65.23±0.02	88.03±0.00	89.11±1.88
DTCCA (Wong et al. 2021)	82.77±0.10	96.01±0.10	80.31±0.22	58.32±0.20	84.01±0.32	90.01±0.26
DUA-Nets (Geng et al. 2021)	87.81±1.44	98.00±0.21	79.83±1.50	65.23±0.11	89.61±1.00	93.01±0.22
MVTCAE (Hwang et al. 2021)	86.21±0.12	97.00±0.23	90.00±0.12	64.22±0.01	93.01±1.22	91.01±0.44
TMC (Han et al. 2021)	88.21±0.41	98.23±0.09	90.04±1.01	66.57±0.03	91.01±0.80	93.01±0.10
TMDOA (Liu et al. 2022)	90.00±0.01	98.30±0.21	91.78±1.22	70.10±0.12	94.01±1.22	93.38±0.04
ETMC (Han et al. 2023)	88.90±0.25	98.20±0.23	90.04±1.21	66.01±0.07	92.02±1.21	92.77±0.37
SMDC (Liu et al. 2023)	93.46±0.00	98.89±0.10	92.06±0.16	72.06±0.22	92.08±1.00	96.40±0.10
ECML (Xu et al. 2024)	92.00±0.03	98.86±0.12	92.67±0.07	73.84±0.19	92.37±0.86	95.33±0.13
RAR (Ours)	**95.65±0.36**	**99.11±0.27**	**95.19±1.59**	**75.69±1.35**	**95.00±0.10**	**97.62±0.37**

Table 2: Comparison with popular multi-view learning methods based on Accuracy (ACC, %).

comprises handwritten numerals ('0'-'9') from Dutch utility maps, represented by six different types of descriptors. **CUB** (Wah et al. 2011) (Caltech-UCSD Birds) contains 11,788 images and text descriptions from 200 bird species, where the first 10 categories are used, with deep features captured from GoogLeNet and text features using doc2vec as two views. **SCENE** (Fei-Fei and Perona 2005) (Scene15) contains 4,485 images from 15 indoor and outdoor scene categories, with GIST, PHOG, and LBP features extracted as multiple views. **MRNet** (Bien et al. 2018) includes knee MRI scans from 1,370 patients, with ACL injury detection as the classification task, using three views. **CAL** (Fei-Fei, Fergus, and Perona 2004) consists of 8,677 images from 101 classes, where the first 10 categories are used, with deep visual features from DECAF and VGG19 as two views.

Implementations: For all datasets except MRNet, we utilized fully connected networks, while MRNet employed ResNet-18 as its backbone. Note that the training dataset is typically clean, so a low value of τ is used during training to enhance classification performance. In the testing phase, various values of τ can be set to control the rejection rate. The Adam optimizer (Kingma and Ba 2014) was used for network training, with l_2-norm regularization set to $1e^{-5}$. A 5-fold cross-validation was employed to select the learning rate from $\{1e^{-5}, 3e^{-4}, 1e^{-3}, 3e^{-3}\}$. For all multi-view datasets, the data was split into training (70%), testing (20%), and calibration (10%) sets. We ran each method 5 times and reported the average values in figures or the mean and standard deviations in tables. The model was implemented in PyTorch and run on a GeForce RTX 4090 GPU with 24GB memory.

Ablation Study

We began by evaluating the impact of each key component of our model: the multi-view fusion strategy, the adaptive rejection module, and the KL-divergence regularization term. In addition to the results in the first row of Table 1, which show the highest accuracy for individual views, we assessed the average performance of our primary components across all views over five runs. As illustrated in Table 1, our method consistently outperforms other combinations, demonstrating the effectiveness of these technical components.

Comparison with Popular Methods

To evaluate the effectiveness of our model, we compared it with several popular multi-view deep classification models. We started by comparing our model with five state-of-the-art evidence-based multi-view methods: TMC (Han et al. 2021), TMDOA (Liu et al. 2022), ETMC (Han et al. 2023), SMDC (Liu et al. 2023), and ECML (Xu et al. 2024). These methods share the same neural network architecture as ours. Given the widespread use of CCA-based methods in multi-view learning, we also compared our approach with three representative CCA-based methods combined with deep learning: DCCA (Andrew et al. 2013), DCCAE (Wang et al. 2015), and DTCCA (Wong et al. 2021). These methods use CCA to derive latent representations, which are then classified using a support vector classifier (SVC). Finally, we compared our model with three advanced multi-view representation learning approaches: CPM-Nets (Zhang et al. 2020), DUA-Nets (Geng et al. 2021), and MVTCAE (Hwang et al. 2021). As in (Hwang et al. 2021), logistic regression was used as the base classifier in these methods,

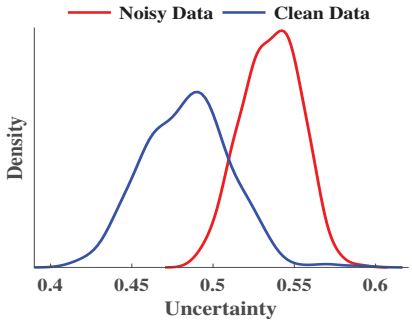

Figure 2: Investigation of our model in capturing data noise. The blue and red curves correspond to the uncertainty density distributions of clean and noisy samples on the HAND dataset, respectively.

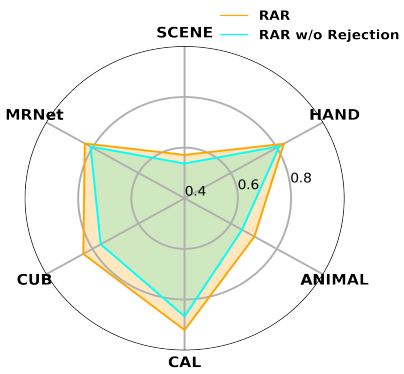

Figure 3: Classification performance with the fusion of noisy views across all datasets. The vertices of the orange line denote the accuracy of multi-view results using RAR, while the vertices of the blue line represent the accuracy of results using RAR without the adaptive rejection mechanism.

with the learned representations as input.

As shown in Table 2, our approach achieves state-of-the-art performance across all multi-view datasets, as reflected in the ACC metric. The comparison among these methods, especially the evidence-based multi-view approaches, underscores the effectiveness of our proposed method. It is important to note that *while the improvement in classification accuracy is notable, it is not the sole objective of RAR*. Instead, RAR is designed to address the challenges associated with non-standard uncertainty measurement in multi-view classification across varying data distributions, thereby enhancing the reliability of multi-view decisions. The subsequent experimental results further support this objective.

Why RAR can Enhance Reliability in Multi-view

In this section, we delve into the practical significance of our method in real-world settings.

Performance Evaluation with Simulated Noises: We assess the reliability of RAR in the presence of noisy samples. We introduce Gaussian noise $\boldsymbol{\epsilon}$ to our samples, transforming them into *noisy* samples: $\widetilde{\boldsymbol{x}}_{\boldsymbol{i}}^{\boldsymbol{v}} = \boldsymbol{x}_i^v + \eta\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is drawn from a Gaussian distribution and η represents the noise intensity. We conducted the following experiments to demonstrate the reliability of our method under these conditions.

(1) Ability to Capture Uncertainty: We first selected one view as the noisy view and added noise with $\eta = 4$ to half of the samples for this view. The density of uncertainty estimated in one view of the HAND dataset is depicted in Figure 2. It can be observed that, for this view, the magnitude of the uncertainty ranges from 0.4 to 0.6. Although the estimated uncertainty density of the noisy samples shows a slight shift compared to that of the clean samples, validating our method's ability to capture noise-induced uncertainty, the small difference in uncertainty levels makes it difficult for users to determine whether to trust the results. Therefore, it is necessary to introduce the adaptive rejection mechanism into the multi-view fusion process.

(2) Enhancing Classification Performance: To evaluate the effectiveness of adaptive rejection in the presence of noisy views, we selected one view as the noisy view and added noise with $\eta = 4$ to all samples in this view across all datasets. Figure 3 presents a radar chart comparing the accuracy of multi-view classification results between RAR and fusion without adaptive rejection. The results show that multi-view classification performance is degraded when noisy views are included, compared to the performance achieved by RAR on noise-free data, as shown in Table 1. However, RAR's classification performance consistently outperforms "RAR w/o Rejection" across all datasets, indicating that the adaptive rejection mechanism effectively mitigates the impact of noisy views and enhances classification performance in noisy environments, thereby improving reliability in multi-view fusion.

(3) Parameter Analysis: We also provide an overall analysis of the parameters used, specifically the effect of noise intensity and the selection of τ. We selected the HAND dataset for this analysis because it has six views (V1-V6), providing more comprehensive information. First, we analyzed the effect of different noise intensities. The results, shown in the first row of Figure 4, demonstrate that the method with adaptive rejection is effective across different noise intensities. We observed that the improvement of RAR increases as η goes from 1 to 3, but then decreases as η increases further from 3 to 10. This is because the magnitude of uncertainty is limited by the specific data distribution; even as noise intensity increases, the uncertainty levels of different samples reach a bound, leading to an initial growth followed by a decline in improvements. Next, we analyzed the effect of τ. The second row of Figure 4 shows that RAR performs better with smaller τ values, which allows for better control of reject risk as τ decreases.

Real-world Application Analysis: Here, we present real-world cases to demonstrate the reliability of our method. Our experimental dataset includes MRNet, a noisy medical multi-view dataset with three views. It contains many noisy samples, as shown in Figure 5. For instance, Cases 1186 and 1194 are filled with noise across all views, and Case 1212 contains artifacts and severe effusion in one or more views, leading to risky predictions. Although these samples exhibit

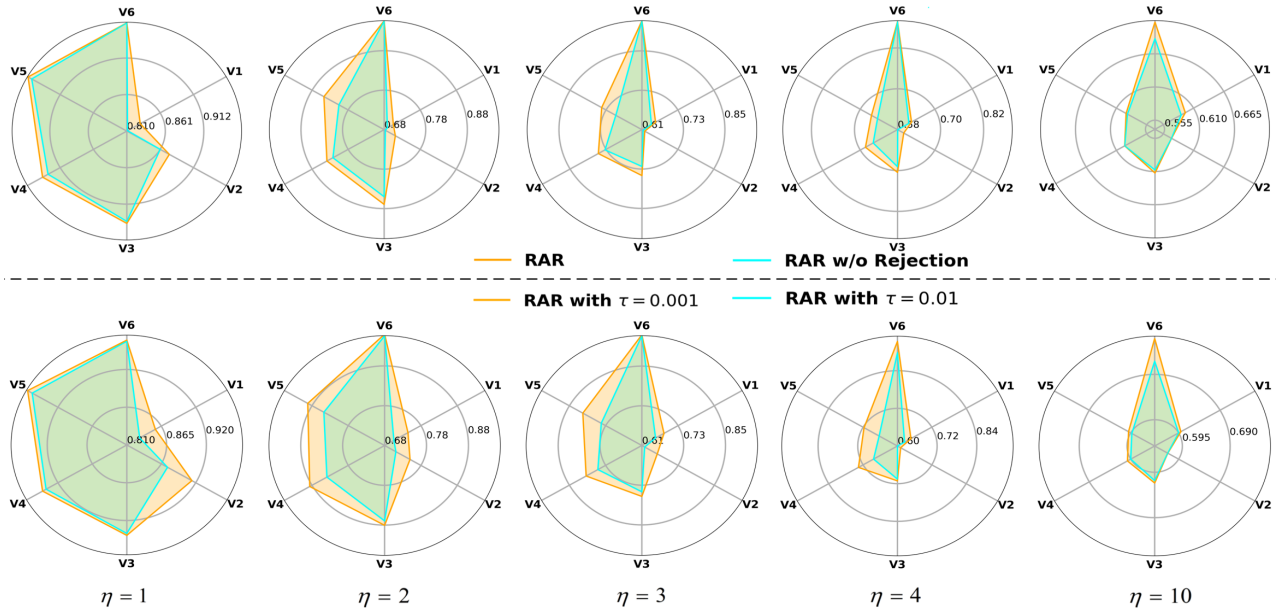

Figure 4: Analysis of Parameters on the HAND dataset with 6 views (V1-V6). The vertices of the line denote the accuracy of multi-view results after integrating the corresponding noisy view with the other 5 views (e.g., V1 indicates noise added to V1). The first row presents a comparison between RAR with $\tau = 0.01$ and RAR without rejection under different noise intensities η. The second row shows the comparison between RAR with $\tau = 0.01$ and RAR with $\tau = 0.001$ under varying noise intensities.

relatively low uncertainty compared to normal samples, as shown in Figure 5, the uncertainty magnitude is small, ranging from 0 to 0.12, making it difficult for doctors to determine whether they should trust the diagnosis. In contrast, our method provides an adaptive rejection threshold u_λ to assess the reliability of predictions, ensuring that unreliable results are flagged and passed to a human doctor for further diagnosis. This process mitigates diagnostic risk and enhances multi-view reliability.

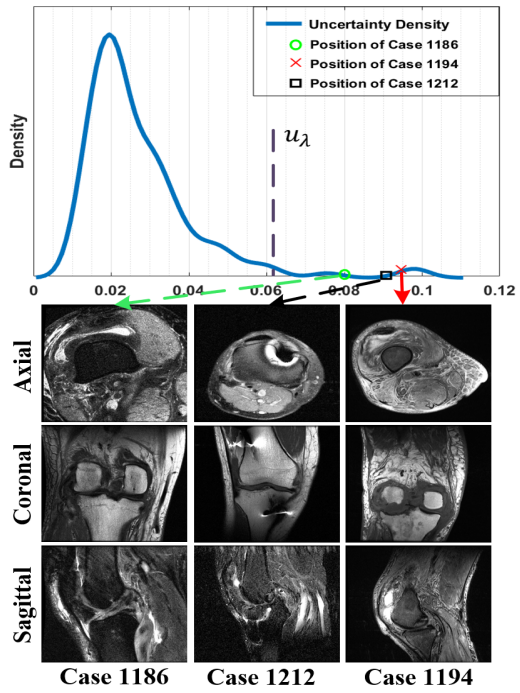

Figure 5: Cases with low-quality views on MRNet. The presented cases are rejected by RAR during the automated decision process and are flagged for further diagnosis by a human doctor.

Conclusion

In this study, we introduced a novel multi-view classification method that integrates adaptive rejection into the fusion process, aiming to improve classification performance and provide clear guidance on whether decision-makers should trust predictions made from low-quality samples, thereby enhancing reliability in multi-view classification. By framing the rejection problem as a multiple-testing issue, we developed an effective adaptive rejection mechanism based on calculated p-values and devised a corresponding fusion strategy. The effectiveness of our approach is demonstrated both theoretically and empirically. Theoretically, we ensure that the false rejection risk is controlled with a user-specified probability. Empirical studies highlight the superior performance of our method across various multi-view datasets, indicating that it effectively mitigates the impact of low-quality views.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62173252, 62476165).

References

Andrew, G.; Arora, R.; Bilmes, J.; and Livescu, K. 2013. Deep canonical correlation analysis. In *International conference on machine learning*, 1247–1255. PMLR.

Angelopoulos, A. N.; Bates, S.; Candès, E. J.; Jordan, M. I.; and Lei, L. 2021. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*.

Benjamini, Y.; and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1): 289–300.

Bien, N.; Rajpurkar, P.; Ball, R. L.; Irvin, J.; Park, A.; Jones, E.; Bereket, M.; Patel, B. N.; Yeom, K. W.; Shpanskaya, K.; et al. 2018. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS medicine*, 15(11): e1002699.

Burgin, M. 2002. The essence of information: Paradoxes, contradictions, and solutions. In *Electronic Conference on Foundations of Information Science: The nature of information: Conceptions, misconceptions, and paradoxes (FIS 2002). Retrieved September*, volume 13, 2013. Citeseer.

De Stefano, C.; Sansone, C.; and Vento, M. 2000. To reject or not to reject: that is the question-an answer in case of neural classifiers. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30(1): 84–94.

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

Fei-Fei, L.; and Perona, P. 2005. 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, 524–531. IEEE.

Franc, V.; Prusa, D.; and Voracek, V. 2023. Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24(11): 1–49.

Geifman, Y.; and El-Yaniv, R. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

Geng, Y.; Han, Z.; Zhang, C.; and Hu, Q. 2021. Uncertainty-Aware Multi-View Representation Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 7545–7553.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2021. Trusted Multi-View Classification. In *International Conference on Learning Representations*.

Han, Z.; Zhang, C.; Fu, H.; and Zhou, J. T. 2023. Trusted Multi-View Classification With Dynamic Evidential Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 2551–2566.

Hwang, H.; Kim, G.-H.; Hong, S.; and Kim, K.-E. 2021. Multi-View Representation Learning via Total Correlation Objective. *Advances in Neural Information Processing Systems*, 34.

Jiang, B.; Wu, X.; Zhou, X.; Cohn, A. G.; Liu, Y.; Sheng, W.; and Chen, H. 2024. Semi-Supervised Multi-View Feature Selection with Adaptive Graph Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(3): 3615–3629.

Jøsang, A. 2018. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Liang, X.; Fu, P.; Guo, Q.; Zheng, K.; and Qian, Y. 2024. DC-NAS: Divide-and-Conquer Neural Architecture Search for Multi-Modal Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 13754–13762.

Liang, X.; Qian, Y.; Guo, Q.; Cheng, H.; and Liang, J. 2021. AF: An association-based fusion method for multi-modal classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 9236–9254.

Liu, W.; Chen, Y.; and Yue, X. 2024. Building Trust in Decision with Conformalized Multi-view Deep Classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7278–7287.

Liu, W.; Chen, Y.; Yue, X.; Zhang, C.; and Xie, S. 2023. Safe Multi-View Deep Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8870–8878.

Liu, W.; Yue, X.; Chen, Y.; and Denoeux, T. 2022. Trusted multi-view deep learning with opinion aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 7585–7593.

Liu, Y.; Liu, L.; Xu, C.; Song, X.; Guan, Z.; and Zhao, W. 2024. Dynamic evidence decoupling for trusted multi-view learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7269–7277.

Lu, J.; Du, L.; Buntine, W.; Jung, M. C.; Dipnall, J.; and Gabbe, B. 2024. Navigating Conflicting Views: Harnessing Trust for Learning. *arXiv preprint arXiv:2406.00958*.

Nouretdinov, I.; Vovk, V.; Vyugin, M.; and Gammerman, A. 2001. Pattern recognition and density estimation under the general iid assumption. In *Computational Learning Theory: 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001 Amsterdam, The Netherlands, July 16–19, 2001 Proceedings 14*, 337–353. Springer.

Patrikar, J.; Veer, S.; Sharma, A.; Pavone, M.; and Scherer, S. 2024. RuleFuser: Injecting Rules in Evidential Networks for Robust Out-of-Distribution Trajectory Prediction. *arXiv preprint arXiv:2405.11139*.

Şensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*.

Shafer, G. 1976. *A mathematical theory of evidence*. Princeton university press.

Sun, S.; Dong, W.; and Liu, Q. 2020. Multi-view representation learning with deep gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

van Breukelen, M.; Duin, R. P.; Tax, D. M.; and Den Hartog, J. 1998. Handwritten digit recognition by combined classifiers. *Kybernetika*, 34(4): 381–386.

Varshney, K. R. 2011. A risk bound for ensemble classification with a reject option. In *2011 IEEE statistical signal processing workshop (SSP)*, 769–772. IEEE.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*.

Wang, M.; Lin, T.; Wang, L.; Lin, A.; Zou, K.; Xu, X.; Zhou, Y.; Peng, Y.; Meng, Q.; Qian, Y.; et al. 2023. Uncertainty-inspired open set learning for retinal anomaly identification. *Nature Communications*, 14(1): 6757.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. 2015. On deep multi-view representation learning. In *International conference on machine learning*, 1083–1092. PMLR.

Wong, H. S.; Wang, L.; Chan, R.; and Zeng, T. 2021. Deep tensor CCA for multi-view learning. *IEEE Transactions on Big Data*, 8(6): 1664–1677.

Xu, C.; Si, J.; Guan, Z.; Zhao, W.; Wu, Y.; and Gao, X. 2024. Reliable conflictive multi-view learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16129–16137.

Xu, Z.; Yue, X.; Lv, Y.; Liu, W.; and Li, Z. 2023. Trusted fine-grained image classification through hierarchical evidence fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10657–10665.

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

Zhang, C.; Zhu, X.; Wang, Z.; Zhong, Y.; Sheng, W.; Ding, W.; and Jiang, B. 2024a. Discriminative Multi-View Fusion via Adaptive Regression. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(6): 3821–3833.

Zhang, L.; Liu, W.; Chen, Y.; and Yue, X. 2022. Reliable multi-view deep patent classification. *Mathematics*, 10(23): 4545.

Zhang, Q.; Wei, Y.; Han, Z.; Fu, H.; Peng, X.; Deng, C.; Hu, Q.; Xu, C.; Wen, J.; Hu, D.; et al. 2024b. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*.

Zhang, X.-Y.; Xie, G.-S.; Li, X.; Mei, T.; and Liu, C.-L. 2023. A survey on learning to reject. *Proceedings of the IEEE*, 111(2): 185–215.

Zhou, H.; Xue, Z.; Liu, Y.; Li, B.; Du, J.; Liang, M.; and Qi, Y. 2023. CALM: An Enhanced Encoding and Confidence Evaluating Framework for Trustworthy Multi-view Learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3108–3116.

Zou, X.; Tang, C.; Zheng, X.; Li, Z.; He, X.; An, S.; and Liu, X. 2023. DPNET: Dynamic Poly-attention Network for Trustworthy Multi-modal Classification. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3550–3559.