Probabilistic Multimodal Learning with von Mises-Fisher Distributions

Peng Hu[1], **Yang Qin**[1], **Yuanbiao Gou**[1], **Yunfan Li**[1], **Mouxing Yang**[1] and **Xi Peng**[1,2]*
[1]College of Computer Science, Sichuan University, China.
[2]National Key Laboratory of Fundamental Algorithms and Models for Engineering Simulation, China.
{penghu.ml, qinyang.gm, gouyuanbiao, yunfanli.gm, yangmouxing, pengx.gm}@gmail.com

Abstract

Multimodal learning is pivotal for the advancement of artificial intelligence, enabling machines to integrate complementary information from diverse data sources for holistic perception and understanding. Despite significant progress, existing methods struggle with challenges such as noisy inputs, noisy correspondence, and the inherent uncertainty of multimodal data, limiting their reliability and robustness. To address these issues, this paper presents a novel Probabilistic Multimodal Learning framework (PML) that models each data point as a von Mises-Fisher (vMF) distribution, effectively capturing intrinsic uncertainty and enabling robust fusion. Unlike traditional Gaussian-based models, PML learns directional representation with a concentration parameter to quantify reliability directly, enhancing stability and interpretability. To enhance discrimination, we propose a von Mises-Fisher Prototypical Contrastive Learning paradigm (vMF-PCL), which projects data onto a hypersphere by pulling within-class samples closer to their class prototype while pushing between-class prototypes apart, adaptively learning the reliability estimations. Building upon the estimated reliability, we develop a Reliable Multimodal Fusion mechanism (RMF) that dynamically adjusts the contribution and conflict of each modality, ensuring robustness against noisy data, noisy correspondence, and uncertainty. Extensive experiments on nine benchmarks demonstrate the superiority of PML, consistently outperforming 14 state-of-the-art methods. Code is available at https://github.com/XLearning-SCU/2025-IJCAI-PML.

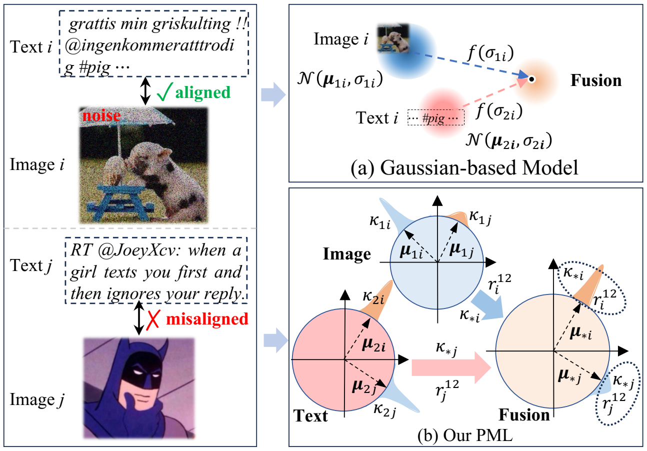


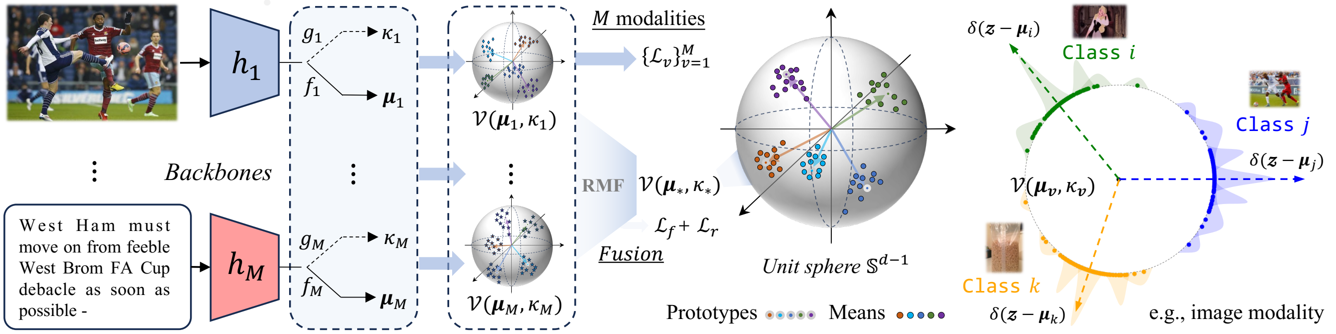
Figure 1: Difference between Gaussian-based models and our PML in handling unreliable multimodal data with noisy inputs and noisy correspondence. Gaussian-based models require a function $f(\cdot)$ to transform estimated variances into reliability for multimodal fusion, often resulting in instability [Li *et al.*, 2021]. In contrast, our PML directly leverages concentration parameters to dynamically adjust the fusion process, which is more stable and reliable. Furthermore, traditional Gaussian-based models focus solely on aleatoric uncertainty, neglecting epistemic uncertainty, which might lead to suboptimal performance. In contrast, our PML could simultaneously handle epistemic and aleatoric uncertainties, enhancing both performance and robustness.

1 Introduction

Multimodal learning integrates consistent and complementary information from diverse data sources to enable a comprehensive perception and understanding of the real world, which is crucial for promoting the intelligence of unmanned systems. Much like humans utilize various sensory organs to perceive the world, machines process and fuse multimodal data from sensors such as cameras, radars, and ultrasonic systems to achieve holistic perception and understanding. By leveraging the complementary strengths of distinct modalities, multimodal learning has significantly advanced applications such as multimodal classification [Xu *et al.*, 2024a; Geng *et al.*, 2021], audio-visual recognition [Afouras *et al.*, 2018], and multi-view clustering [Yang *et al.*, 2022; Wen *et al.*, 2023b]. The core of multimodal learning lies in effectively extracting and integrating consistent and complementary information from different modalities to enhance performance for various tasks.

To integrate various modalities, numerous multimodal learning techniques have been developed, ranging from early fusion [Yu *et al.*, 2021], which concatenates features at the

*Corresponding author.


h_1
h_M
Backbones
West Ham must move on from feeble West Brom FA Cup debacle as soon as possible -
g_1
κ_1
f_1
μ_1
g_M
κ_M
f_M
μ_M
$\mathcal{V}(\mu_1, \kappa_1)$
$\mathcal{V}(\mu_M, \kappa_M)$
M modalities
$\{\mathcal{L}_v\}_{v=1}^M$
RMF
$\mathcal{V}(\mu_*, \kappa_*)$
$\mathcal{L}_f + \mathcal{L}_r$
Fusion
Unit sphere $\mathbb{S}^{d-1}$
Prototypes
Means
$\delta(z - \mu_i)$
Class i
Class j
$\delta(z - \mu_j)$
$\mathcal{V}(\mu_v, \kappa_v)$
Class k
$\delta(z - \mu_k)$
e.g., image modality

(a) vMF-based Prototypical Contrastive Learning

(b) vMF Representation

Figure 2: Overview of our Probabilistic Multimodal Learning (PML) framework. First, PML utilizes modality-specific backbones to project the data into a latent space. Then, PML projects each point into a distribution described by a mean direction ($\boldsymbol{\mu}$) and a concentration parameter (κ), enabling intrinsic uncertainty estimation. To obtain discriminative directional representations, PML exploits vMF-based prototype contrastive learning to maximize the agreement between the data and the corresponding class prototype in the latent hypersphere.

input or feature level, to late fusion [Cao *et al.*, 2024], which integrates predictions at the decision level. Although these approaches have achieved promising results, they often face significant challenges arising from modality imbalance [Peng *et al.*, 2022], noise [Cao *et al.*, 2024], and incomplete or uncertain data [Xie *et al.*, 2023; Xu *et al.*, 2024a]. For instance, some modalities might dominate others during fusion, causing the model to overlook weak but informative signals. In addition, noise in one modality or misalignment between modalities (*a.k.a.* noisy correspondence) would remarkably degrade performance. Although current efforts have introduced robust fusion mechanisms and cross-modal attention to mitigate these issues [Cao *et al.*, 2024], the aleatoric and epistemic uncertainty [Kendall and Gal, 2017] inherent in multimodal data and model remains underexplored. The uncertainty arises from both aleatoric factors (intrinsic noise in data) and epistemic factors (limited knowledge or model capacity), posing a fundamental limitation to the reliability of existing multimodal learning systems [Gao *et al.*, 2024].

To address these challenges, reliable multimodal learning has emerged as a critical research direction, focusing on building models capable of managing uncertainty and ensuring robustness [Geng *et al.*, 2021; Xie *et al.*, 2023; Xu *et al.*, 2024a]. By explicitly quantifying and incorporating uncertainty, these methods aim to enhance trustworthiness and interpretability, enabling more reliable decision-making in safety-critical applications, such as autonomous driving and medical diagnosis [Tang *et al.*, 2022; Zou *et al.*, 2024]. Techniques like Bayesian learning [Kendall and Gal, 2017], ensemble methods [Lakshminarayanan *et al.*, 2017], and uncertainty-aware attention mechanisms [Heo *et al.*, 2018] have shown promising results, but their practical applications often encounter significant computational burdens or strong assumptions, which limit their scalability and generalizability. As a remedy, Evidential Deep Learning (EDL) [Sensoy *et al.*, 2018; Xu *et al.*, 2024a] offers an alternative by treating predictions as subjective opinions and directly inferring uncertainty. However, existing methods struggle to distinguish between epistemic and aleatoric uncertainty, limiting their capability to handle open scenarios. For example, noisy data would produce aleatoric uncertainty, while noisy correspondence would induce epistemic uncertainty as shown in Figure 1.

In this work, we propose a novel Probabilistic Multimodal Learning framework (PML), as shown in Figure 2, which models each multimodal data point as a von Mises-Fisher (vMF) distribution for reliable fusion. Unlike Gaussian-based models, which rely on variance to capture uncertainty, the vMF distribution represents data as directional distributions with a concentration parameter that directly quantifies reliability, thus embracing better stability. Specifically, our PML first projects each data point into a distribution described by a mean direction and a concentration parameter, enabling intrinsic uncertainty estimation within the data. We then present a von Mises-Fisher Prototypical Contrastive Learning paradigm (vMF-PL) to pull within-class samples closer to their category prototype while pushing between-class prototypes apart on the hypersphere, inherently capturing the directional discrimination and reliability in the multimodal data. By leveraging the estimated reliability, a Reliable Multimodal Fusion mechanism (RMF) is proposed to dynamically weight mean directions of different modalities for reliable classification, enhancing robustness to data noise, cross-modal misalignment, and uncertainty. Moreover, cross-entropy is utilized to evaluate the epistemic reliability between the probabilistic predictions of each modality and fusion results. In the inference stage, the learned aleatoric and epistemic reliability values are used to fuse the predictions for reliable and robust classification. Extensive experiments on nine diverse benchmarks validate the effectiveness and reliability of our PML, demonstrating its superiority over existing methods.

The major contributions of this work can be summarized below:

- We propose a novel vMF-based Probabilistic Multimodal Learning framework (PML), which captures intrinsic aleatoric and epistemic reliability to dynamically mitigate unreliable modalities, facilitating stable and robust classification.

- To help capture both directional information and reliability of each modality, we present a von Mises-Fisher Prototypical Contrastive Learning paradigm (vMF-PCL), which enhances within-class compactness and between-class scatterness.
- Extensive experiments on nine widely-used multimodal benchmarks demonstrate the effectiveness and robustness of the proposed PML, outperforming 14 state-of-the-art baselines.

2 Related Work

In this section, we briefly review two key areas most relevant to this paper: multimodal learning and uncertainty-aware learning.

2.1 Multimodal Learning

Multimodal learning approaches can be broadly categorized into three types: early fusion, middle fusion, and late fusion, depending on the stage at which information aggregation occurs. Early fusion methods commonly combine multimodal data at the input level [Yu *et al.*, 2021], allowing direct interactions between modalities. However, these methods often struggle with scalability and modality-specific noise. Middle fusion methods [Natarajan *et al.*, 2012] integrate multimodal data at the feature level by fusing intermediate representations. Unlike these methods, late fusion approaches aggregate predictions at the decision level [Cao *et al.*, 2024], offering greater flexibility while sacrificing the exploitation of inter-modal relationships. Recent advancements focus on more sophisticated fusion techniques, such as attention mechanisms [Nagrani *et al.*, 2021], cross-modal interactions [Chen *et al.*, 2019], and graph-based methods [Mai *et al.*, 2020]. These approaches aim to address challenges such as modality imbalance [Peng *et al.*, 2022] and noisy inputs [Wen *et al.*, 2023a; Zhang *et al.*, 2023; Cao *et al.*, 2024] by dynamically adjusting contributions from each modality. Despite their potential, most methods do not explicitly account for uncertainty, limiting their robustness and reliability in real-world scenarios.

2.2 Uncertainty-aware Learning

Uncertainty modeling has gained significant attention as a critical aspect of reliable AI systems, particularly in safety-critical applications. Epistemic uncertainty, arising from model limitations, and aleatoric uncertainty, stemming from intrinsic data noise, are two primary types of uncertainty [Kendall and Gal, 2017]. Various deep learning techniques have been proposed to quantify and leverage uncertainty, including Bayesian neural networks [Kendall and Gal, 2017], Monte Carlo sampling [Zhang, 2021], and ensemble learning [Lakshminarayanan *et al.*, 2017]. Recently, in the context of multimodal learning, robust methods incorporating uncertainty have emerged, such as uncertainty-aware attention [Heo *et al.*, 2018], probabilistic embeddings (PE) [Gao *et al.*, 2024; Shi and Jain, 2019] and evidential deep learning (EDL) [Sensoy *et al.*, 2018; Xu *et al.*, 2024a]. Most existing techniques focus on modeling only one type of uncertainty, often neglecting the potential benefits of jointly quantifying both epistemic and aleatoric uncertainties. While EDL provides a direct estimation of uncertainty through subjective opinions, it struggles to disentangle the two types of uncertainties. Our work fulfills this gap by leveraging the concentration parameter of the vMF distribution to directly quantify aleatoric reliability, and the cross-entropy across probabilistic predictions to estimate epistemic reliability, thus enabling more comprehensive uncertainty modeling in multimodal learning.

3 Methodology

In this section, we introduce the proposed Probabilistic Multimodal Learning framework (PML), which leverages the von Mises-Fisher (vMF) distribution for reliable multimodal fusion and classification. Our PML consists of three key components: (1) vMF-based feature representation, (2) vMF-based Prototypical Contrastive Learning (vMF-PCL), and (3) Reliable Multimodal Fusion (RMF). Each component is elaborated below.

3.1 vMF-based Feature Representation

For ease of representation, we first give definitions for multimodal classification. Let $\mathcal{D} = \{\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \cdots, \boldsymbol{x}_{Mi}, y_i\}_{i=1}^{N}$ denote the multimodal training set, where $\boldsymbol{x}_{vi}$ represents the v-th modality of the i-th instance, $y_i \in \{1, 2, \cdots, K\}$ denotes the corresponding ground-truth label, K is the number of classes, M is the number of modalities, and N is the number of instances. Multimodal learning aims to effectively utilize information from each modality to achieve comprehensive perception and understanding. However, there is inevitable noise in the multimodal inputs, such as corrupted data and noisy data, resulting in aleatoric uncertainty.

To capture both directional representation and intrinsic aleatoric uncertainty, we model each unimodal sample $\boldsymbol{x}_{vi}$ as a vMF distribution $\boldsymbol{z} \sim \mathcal{V}_d(\boldsymbol{\mu}_{vi}, \kappa_{vi})$, where $\boldsymbol{\mu}_{vi}$ and κ_{vi} denote the mean direction and concentration parameter, respectively. The larger value of κ_{vi} refers to the higher concentration of the distribution around the mean direction $\boldsymbol{\mu}_{vi}$, as well as the reliability. Specifically, given an input sample $\boldsymbol{x}_{vi}$ from the v-th modality, two modality-specific sub-networks are exploited to estimate the mean directional representation $\boldsymbol{\mu}_{vi}$ and the concentration parameter κ_{vi} as follows:

$$\boldsymbol{\mu}_{vi} = f_v(h_v(\boldsymbol{x}_{vi})) \in \mathbb{R}^d, \kappa_{vi} = g_v(h_v(\boldsymbol{x}_{vi})) \in \mathbb{R}^1, \quad (1)$$

where d is the dimensionality of the latent space, h_v refers to the backbone for the v-th modality, and $f_v(\cdot)$, $g_v(\cdot)$ are two different sub-networks used to estimate the mean direction and the concentration parameter respectively. Then, each point $\boldsymbol{x}_{vi}$ could be modeled by a specific vMF distribution defined on a d-dimensional unit sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$ [Li *et al.*, 2021; Conti *et al.*, 2022], with the following probability density function:

$$p(\boldsymbol{z}|\boldsymbol{\mu}_{vi}, \kappa_{vi}) = C_d(\kappa_{vi}) \exp(\kappa_{vi}\boldsymbol{\mu}_{vi}^{\top}\boldsymbol{z}), \quad (2)$$

$$C_d(\kappa_{vi}) = \frac{\kappa_{vi}^{d/2-1}}{(2\pi)^{d/2}\mathcal{I}_{d/2-1}(\kappa_{vi})}, \quad (3)$$

where $C_d(\kappa_{vi})$ is the normalization constant dependent on the dimensionality d of $\boldsymbol{z}$, and $\mathcal{I}_\nu(\cdot)$ is the modified Bessel function of the first kind, defined as:

$$\mathcal{I}_\nu(t) = \sum_{m=0}^{\infty} \frac{\left(\frac{t}{2}\right)^{2m+\nu}}{m!\Gamma(m+\nu+1)}. \tag{4}$$

The concentration parameter κ_{vi} quantifies the reliability of the sample $\boldsymbol{x}_{vi}$, with larger values indicating higher certainty.

3.2 vMF-based Prototypical Contrastive Learning

In the ideal scenario, a well-trained model should group within-class samples together while separating between-class samples apart, thereby enhancing the representation discriminability. In other words, each class could be ideally represented as a spherical Dirac delta distribution $\delta(\boldsymbol{z} - \bar{\boldsymbol{\mu}}_{vk})$, which is a degenerate distribution to concentrate all mass at a single point $\bar{\boldsymbol{\mu}}_{vk}$, *i.e.*, the unit vector of the k-th class from the v-th modality. Here, $\delta(\boldsymbol{z} - \bar{\boldsymbol{\mu}}_{vk})$ represents the features of the k-th class from the v-th modality, serving as the k-th class prototype. Intuitively, it is reasonable to assume that the samples of the same class are distributed around their corresponding category prototype $\delta(\boldsymbol{z} - \bar{\boldsymbol{\mu}}_{vk})$ on the latent hypersphere. Let $\mathcal{C}_v = \{\delta(\boldsymbol{z} - \bar{\boldsymbol{\mu}}_{vk})\}_{k=1}^{K}$ represent the prototypes of K classes for the v-th modality. Formally, for the given prototype $\delta(\boldsymbol{z} - \bar{\boldsymbol{\mu}}_{vk})$ of the k-th class on the hypersphere, the difference between a point $\mathcal{V}_d(\boldsymbol{\mu}_{vi}, \kappa_{vi})$ and a prototype $\delta(\boldsymbol{z} - \bar{\boldsymbol{\mu}}_{vk})$ could be computed as follows:

$$\begin{aligned} l_{ik}^{v} &= D_{KL}\left(\delta(\boldsymbol{z} - \bar{\boldsymbol{\mu}}_{vk}) \| \mathcal{V}_d(\boldsymbol{\mu}_{vi}, \kappa_{vi})\right) \\ &= -\kappa_{vi}\boldsymbol{\mu}_{vi}^{\top}\bar{\boldsymbol{\mu}}_{vk} + \log(\mathcal{I}_{d/2-1}(\kappa_{vi})) \\ &\quad - (\frac{d}{2} - 1)\log(\kappa_{vi}) + \frac{d}{2}\log(2\pi). \end{aligned} \tag{5}$$

Then, the probability p_{ij}^{v} of a point $\boldsymbol{x}_{vi}$ belonging to the j-th class could be computed as:

$$p_{ij}^{v} = \sigma\left(-[l_{i1}^{v}, l_{i2}^{v}, \cdots, l_{iK}^{v}]\right)_j = \frac{\exp(-l_{ij}^{v})}{\sum_{k=1}^{K}\exp(-l_{ik}^{v})}, \tag{6}$$

where σ represents the *softmax* function. Ideally, if the point $\boldsymbol{x}_{vi}$ belongs to the j-th class, p_{ij}^{v} should be maximized; otherwise, it should be minimized. Following contrastive learning principles [Chen *et al.*, 2020], we formulate the following loss function by using the negative log-likelihood:

$$\mathcal{L}_v = -\frac{1}{N}\sum_{i=1}^{N}\ell(\boldsymbol{\mu}_{vi}, \kappa_{vi}, y_i), \tag{7}$$

$$\ell(\boldsymbol{\mu}_{vi}, \kappa_{vi}, y_i) = \log\frac{\exp(-l_{iy_i}^{v})}{\sum_{k=1}^{K}\exp(-l_{ik}^{v})}. \tag{8}$$

By minimizing this loss function, we could maximize the agreement between the data and their corresponding class prototypes in the latent hypersphere, thus encapsulating the discrimination into directional representations. Then, we could obtain the overall loss for all modalities as:

$$\mathcal{L}_{vMF} = \sum_{v=1}^{M}\mathcal{L}_v. \tag{9}$$

3.3 Reliable Multimodal Fusion

Multimodal fusion aims to learn m modality-specific transformations $f_v(\cdot)$, $h_v(\cdot)$ to project each modality $\mathcal{X}_v = \{\boldsymbol{x}_{vi}\}_{i=1}^{N}$ into feature representations $\mathcal{Z}_v = \{\boldsymbol{z}_{vi}\}_{i=1}^{N}$, which are then integrated for comprehensive decision as follows:

$$\boldsymbol{q}_i = \phi\left(f_1(h_1(\boldsymbol{x}_{1i})), \cdots, f_M(h_M(\boldsymbol{x}_{Mi}))\right), \tag{10}$$

where $\phi(\cdot)$ is a fusion function that leverages both consistent and complementary information for holistic predictions. For the early fusion, $\phi(\cdot)$ is a nonlinear function, while $f(\cdot)$ could be a linear or nonlinear function. On the other hand, for the late fusion, $f(\cdot)$ is a nonlinear transformation and $\phi(\cdot)$ is a linear transformation. In this paper, we focus on late fusion. To simplify the presentation, the widely-used *softmax* function is employed to compute the probability of $\boldsymbol{x}_{vi}$ belonging to the j-th class, namely:

$$\begin{aligned} \boldsymbol{q}_{ij}^{v} &= \sigma\left(\boldsymbol{W}^{\top}\left[f_1(h_1(\boldsymbol{x}_{1i})), \cdots, f_M(h_M(\boldsymbol{x}_{Mi}))\right]\right)_j \\ &= \frac{\exp\left(\sum_{v=1}^{M}\boldsymbol{w}_{vj}^{\top}f_v(h_v(\boldsymbol{x}_{vi}))\right)}{\sum_{k=1}^{K}\exp\left(\sum_{v=1}^{M}\boldsymbol{w}_{vk}^{\top}f_v(h_v(\boldsymbol{x}_{vi}))\right)}, \end{aligned} \tag{11}$$

where σ is the *softmax* function, $\boldsymbol{w}_{vj}$ is the projection of the j-th category for the v-th modality, $\boldsymbol{W} = [\boldsymbol{w}_{\cdot 1}, \boldsymbol{w}_{\cdot 2}, \cdots, \boldsymbol{w}_{\cdot K}]$ is the projection matrix acting as a classifier for the concatenated representations, and $\boldsymbol{w}_{\cdot k} = [\boldsymbol{w}_{1k}^{\top}, \boldsymbol{w}_{2k}^{\top}, \cdots, \boldsymbol{w}_{vk}^{\top}]^{\top}$. Obviously, the prototypes could be these modality-specific classifiers $\{\boldsymbol{w}_{vj} | v = 1, 2, \cdots, M; j = 1, 2, \cdots, K\}$. Then, Equation (11) could be rewritten as:

$$\boldsymbol{q}_{ij}^{v} = \frac{\exp\left(\sum_{v=1}^{M}\boldsymbol{\mu}_{vi}^{\top}\bar{\boldsymbol{\mu}}_{vj}\right)}{\sum_{k=1}^{K}\exp\left(\sum_{v=1}^{M}\boldsymbol{\mu}_{vi}^{\top}\bar{\boldsymbol{\mu}}_{vk}\right)}. \tag{12}$$

However, the naive fusion approach treats all modalities equally, making it susceptible to intrinsic uncertainties in the data [Han *et al.*, 2022a; Xu *et al.*, 2024a]. To address this problem, we incorporate the estimated reliability to weight the mean directional representations for more reliable predictions:

$$\begin{aligned} \boldsymbol{q}_{ij}^{v} &= \frac{\exp\left(\sum_{v=1}^{M}\alpha_{vi}\boldsymbol{\mu}_{vi}^{\top}\bar{\boldsymbol{\mu}}_{vj}\right)}{\sum_{k=1}^{K}\exp\left(\sum_{v=1}^{M}\alpha_{vi}\boldsymbol{\mu}_{vi}^{\top}\bar{\boldsymbol{\mu}}_{vk}\right)} \\ &= \sigma\left(\boldsymbol{V}^{\top}\left[\alpha_{1i}\boldsymbol{\mu}_{1i}^{\top}, \alpha_{2i}\boldsymbol{\mu}_{2i}^{\top}, \cdots, \alpha_{Mi}\boldsymbol{\mu}_{Mi}^{\top}, \right]\right)_j, \end{aligned} \tag{13}$$

where $\alpha_{vi} = \frac{\kappa_{vi}}{\sum_{v=1}^{M}\kappa_{vi}}$, $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_K]$ represents the reliability weights for each modality, and $\boldsymbol{v}_k = [\boldsymbol{\mu}_{1k}^{\top}, \boldsymbol{\mu}_{2k}^{\top}, \cdots, \boldsymbol{\mu}_{vk}^{\top}]^{\top}$.

Let $\boldsymbol{\mu}_{*i} = \epsilon[\alpha_{1i}\boldsymbol{\mu}_{1i}, \alpha_{2i}\boldsymbol{\mu}_{2i}, \cdots, \alpha_{Mi}\boldsymbol{\mu}_{Mi}] \in \mathbb{R}^{d\times M}$, which could also be viewed as a directional vector on a hypersphere, where ϵ is a normalization factor ensuring $\|\boldsymbol{\mu}_{*i}\| = 1$. Consequently, the fused representations are modeled as a vMF distribution $\mathcal{V}_{d\times M}(\boldsymbol{\mu}_{*i}, \kappa_{*i})$, where $\kappa_{*i} = \frac{1}{M}\sum_{v=1}^{M}\kappa_{vi}$ is the concentration parameter of the fused vMF distribution, capturing the overall aleatoric reliability of the fused multimodal inputs. Following vMF-PCL, we could utilize the negative log-likelihood to formulate the fusion loss function as follows:

$$\mathcal{L}_f = -\frac{1}{N}\sum_{i=1}^{N} \ell(\boldsymbol{\mu}_{*i}, \kappa_{*i}, y_i). \tag{14}$$

In order to estimate epistemic reliability, the entropy is employed on the probability as follows:

$$r_i^{uv} = \exp\left(\sum_{k=1}^{K} p_{ik}^u \log\left(p_{ik}^v\right)\right), \tag{15}$$

where r_i^{uv} measures the prediction consistency between the u-th and v-th modalities for the i-th point. When $v = 0$, r_i^{u*} denotes the prediction agreement between the v-th modality and its final decision, and when $v = u$, r_i^{uu} indicates the reliability of the u-th modality, particularly for the fused prediction ($u = v = 0$). Intuitively, since epistemic reliability should align with aleatoric reliability within the same modality, we present a reliability consistency regularizer to enhance reliability estimation:

$$\mathcal{L}_r = \frac{1}{N}\sum_{i=1}^{N}\sum_{v=1}^{M} |\alpha_{vi} - \beta_{vi}|, \tag{16}$$

where $\beta_{vi} = \frac{r_i^{vv}}{\sum_{v=1}^{M} r_i^{vv}}$ represents the normalized prediction reliability for the v-th modality. Then, we could obtain the final loss of RMF as:

$$\mathcal{L}_{RMF} = \mathcal{L}_f + \mathcal{L}_r. \tag{17}$$

3.4 Optimization Objective

The overall training objective combines the vMF-PCL losses across all modalities, the fused hypersphere, and a reliability consistency regularizer. The final objective is given by:

$$\mathcal{L} = \mathcal{L}_{vMF} + \mathcal{L}_{RMF}. \tag{18}$$

Our proposed PML could be iteratively trained by minimizing the objective function Equation (18) in a batch-by-batch manner using one stochastic gradient descent optimization algorithm, such as Adam [Kingma and Ba, 2014]. By minimizing $\mathcal{L}$, our model could effectively capture both the discriminability and reliability from the multimodal data.

During inference, the prediction consistency r_i^{v*} and learned reliability κ_{vi} are utilized to weigh the mean directions of each modality:

$$\alpha'_{vi} = \frac{1}{2}\left(\frac{r_i^{v*}}{\sum_{k=1}^{M} r_i^{k*}} + \frac{\kappa_{vi}}{\sum_{k=1}^{M}\kappa_{ki}}\right). \tag{19}$$

Then the obtained weights are applied to fuse the learned directional representations like $[\alpha'_{1i}\boldsymbol{\mu}_{1i}^{\top}, \alpha'_{2i}\boldsymbol{\mu}_{2i}^{\top}, \cdots, \alpha'_{Mi}\boldsymbol{\mu}_{Mi}^{\top},]$ for final classification according to Equation (13).

4 Experiments

In this section, we conduct comprehensive experiments to demonstrate the effectiveness of our PML across nine widely-used benchmarks and compare it with 14 state-of-the-art baselines. Besides, we provide detailed ablation studies to analyze the contributions of individual components.

4.1 Datasets and Experimental Setup

Datasets. We evaluate our framework on nine publicly available benchmark datasets spanning diverse modalities, such as image-text, audio-visual, and feature fusion tasks, to highlight the generalizability of our approach. These datasets comprise **Handwritten**[Duin, 1998], **MSRC-V1**[1], **NUS-WIDE-OBJECT**[2] (NUSOBJ) [Chua *et al.*, 2009], **Fashion-MV** [Xu *et al.*, 2022], **Scene15**[3], **LandUse** [Yang and Newsam, 2010], **Leaves100**[4], **MVSA** [Niu *et al.*, 2016] and **UPMC-Food101** [Wang *et al.*, 2015]. Specifically, the Handwritten consists of six feature sets to characterize 2,000 instances of handwritten digits from "0" to "9", with each category containing 200 instances. The NUSOBJ dataset contains 30,000 instances from 31 categories with five different views. The MSRC-v1 dataset contains 210 images categorized into seven classes, each with five views. The Fashion-MV dataset is a multi-view version of the Fashion dataset [Xiao *et al.*, 2017], designed for multi-view learning analysis, which consists of 30,000 examples of fashion products, divided into ten categories (*e.g.*, T-shirt, Dress, Coat). The Scene15 dataset has 4,485 images categorized into 15 indoor and outdoor scene classes, wherein each image is represented using GIST, PHOG, and LBP. The LandUse dataset contains 2,100 satellite images classified into 21 categories, where each image is represented using three different feature extraction techniques. The Leaves100 dataset contains 1,600 leaf samples from 100 plant species, with three feature views extracted for each leaf image: shape descriptors, fine-scale edges, and texture histograms. The MVSA dataset is a sentiment analysis dataset that consists of more than 2,000 image-text pairs collected from social media. The UPMC-Food101 dataset is a large multimodal food dataset and consists of more than 100,000 recipes for a total of 101 categories.

Experimental setup. In our experiments, in addition to the normal experimental setting, we also construct a noise setting following [Xu *et al.*, 2024a] to further evaluate the robustness of our method by introducing data noise and noisy correspondences (misaligned views/modalities) on test sets. We report the accuracies on the test set to measure the performance. To be further convincing, we report the mean and standard deviation of experiments conducted with 10 random seeds following the format of "mean ± std".

4.2 Implementation Details

In our experiments, the sub-networks f_v and g_v, which estimate the mean direction and concentration parameter, respectively, are designed as multi-layer linear networks. We exploit the Adam optimizer with a batch size of 32 to train all models, using a learning rate of $1e$-4 for all datasets.

4.3 Comparisons with State of the Art

To verify the effectiveness and robustness of our method, we compare our PML against 14 baselines, including

[1] https://mldta.com/dataset/msrc-v1/home/

[2] https://lms.comp.nus.edu.sg/wp-content/uploads/2019/

[3] https://doi.org/10.6084/m9.figshare.7007177.v1

[4] https://archive.ics.uci.edu/dataset/241/one+hundred+plant+species+leaves+data+set

Methods	Handwritten	MSRC-V1	NUSOBJ	Fashion-MV	Scene15	LandUse	Leaves100
DUA-Nets (AAAI’21)	98.10 ± 0.32	84.67 ± 3.03	27.75 ± 0.00	91.08 ± 0.17	65.01 ± 1.55	45.24 ± 1.85	90.31 ± 1.25
TMC (ICLR’21)	98.51 ± 0.15	91.70 ± 2.70	38.77 ± 0.81	95.40 ± 0.40	67.71 ± 0.30	31.69 ± 3.93	86.81 ± 2.20
ETMC (TPAMI’22)	98.75 ± 0.00	92.86 ± 3.01	44.23 ± 0.76	96.21 ± 0.36	71.61 ± 0.28	43.52 ± 3.19	91.44 ± 2.39
TMDL-OA (AAAI’22)	98.55 ± 0.45	95.00 ± 1.67	27.88 ± 0.67	86.52 ± 0.04	75.57 ± 0.02	25.02 ± 2.10	75.28 ± 3.57
DFTMC (CVPR’22)	98.75 ± 0.39	96.90 ± 2.14	-	-	63.10 ± 3.60	34.95 ± 1.69	69.92 ± 2.54
DCP-CV (TPAMI’22)	97.91 ± 0.59	92.86 ± 2.61	32.19 ± 9.48	97.96 ± 0.16	76.70 ± 2.15	71.71 ± 2.09	95.62 ± 1.38
DCP-CG (TPAMI’22)	99.00 ± 0.47	95.24 ± 3.69	43.65 ± 1.10	98.11 ± 0.23	77.79 ± 1.73	75.74 ± 0.98	98.19 ± 0.46
UIMC (CVPR’23)	98.25 ± 0.00	98.81 ± 1.19	43.42 ± 0.12	98.13 ± 0.13	77.70 ± 0.00	57.95 ± 0.61	95.31 ± 0.71
QMF (ICML’23)	98.72 ± 0.48	97.86 ± 1.28	38.13 ± 0.73	98.93 ± 0.32	68.58 ± 1.49	47.86 ± 2.55	95.69 ± 1.25
ECML (AAAI’24)	98.72 ± 0.39	94.05 ± 1.60	39.10 ± 0.74	95.25 ± 0.46	76.19 ± 0.12	60.10 ± 2.01	92.53 ± 1.94
TMNR (IJCAI’24)	97.20 ± 0.63	94.05 ± 3.24	34.52 ± 0.85	94.10 ± 0.50	68.10 ± 1.15	27.38 ± 1.88	90.13 ± 1.53
CCML (MM’24)	97.60 ± 0.62	96.90 ± 2.39	41.43 ± 0.71	95.16 ± 0.41	73.02 ± 1.44	44.86 ± 2.03	97.72 ± 0.92
PDF (ICML’24)	98.40 ± 0.37	97.14 ± 1.78	46.78 ± 0.33	98.95 ± 0.19	70.25 ± 1.21	45.17 ± 2.66	98.03 ± 0.71
PML (Ours)	**99.32 ± 0.45**	**99.52 ± 0.95**	**49.16 ± 0.40**	**99.10 ± 0.22**	**82.70 ± 0.86**	**82.05 ± 1.36**	**99.91 ± 0.14**

Table 1: Accuracy (%) performance on normal test sets. The best and second-best results are in bold and underlined, respectively.

Methods	Handwritten	MSRC-V1	NUSOBJ	Fashion-MV	Scene15	LandUse	Leaves100
DUA-Nets (AAAI’21)	87.16 ± 0.34	78.57 ± 4.45	25.64 ± 0.25	83.03 ± 0.18	26.18 ± 1.31	37.22 ± 0.56	65.62 ± 2.19
TMC (ICLR’21)	92.76 ± 0.15	86.20 ± 4.90	36.00 ± 0.78	84.76 ± 0.78	42.27 ± 1.61	19.67 ± 1.88	70.25 ± 2.55
ETMC (TPAMI’22)	93.85 ± 1.26	87.14 ± 4.54	40.45 ± 0.81	86.48 ± 1.05	56.90 ± 1.70	36.05 ± 2.50	74.19 ± 1.74
TMDL-OA (AAAI’22)	92.45 ± 0.05	84.52 ± 2.20	27.02 ± 0.75	74.55 ± 0.07	48.42 ± 1.02	21.71 ± 1.83	62.28 ± 3.70
DFTMC (CVPR’22)	53.65 ± 20.07	60.24 ± 23.45	-	-	36.01 ± 2.78	7.88 ± 0.94	1.10 ± 0.12
DCP-CV (TPAMI’22)	97.91 ± 0.80	84.76 ± 7.00	28.10 ± 7.80	92.72 ± 2.41	66.22 ± 2.12	59.98 ± 1.93	76.94 ± 1.36
DCP-CG (TPAMI’22)	98.20 ± 0.56	90.00 ± 1.78	38.61 ± 1.29	90.38 ± 2.17	66.44 ± 0.32	61.83 ± 2.48	79.06 ± 1.22
UIMC (CVPR’23)	97.78 ± 0.24	96.90 ± 1.09	41.72 ± 0.31	89.71 ± 0.25	68.25 ± 0.47	50.43 ± 0.46	80.25 ± 1.05
QMF (ICML’23)	97.52 ± 0.86	95.95 ± 1.52	35.62 ± 0.90	92.69 ± 0.78	59.53 ± 1.63	40.17 ± 2.67	77.47 ± 1.46
ECML (AAAI’24)	94.52 ± 0.79	90.00 ± 2.78	36.51 ± 0.76	84.10 ± 0.88	56.97 ± 0.52	50.31 ± 1.81	74.88 ± 1.89
TMNR (IJCAI’24)	92.78 ± 1.01	90.71 ± 4.19	30.88 ± 0.58	85.76 ± 0.81	60.00 ± 1.43	23.95 ± 1.92	74.09 ± 1.99
CCML (MM’24)	93.22 ± 1.09	94.29 ± 2.18	37.38 ± 0.65	83.84 ± 1.01	62.08 ± 1.34	37.76 ± 1.93	78.87 ± 2.31
PDF (ICML’24)	94.35 ± 1.21	94.52 ± 3.02	43.57 ± 0.36	90.73 ± 0.53	58.75 ± 1.03	39.40 ± 1.94	76.34 ± 1.26
PML (Ours)	**98.77 ± 0.48**	**97.86 ± 2.49**	**46.95 ± 0.50**	**96.33 ± 0.21**	**72.63 ± 1.28**	**71.93 ± 2.14**	**89.69 ± 1.50**

Table 2: Accuracy (%) performance on noisy test sets. The best and second-best results are in bold and underlined, respectively.

Methods	MVSA	UPMC-Food101
CONCATENATION	70.71 ± 3.08	88.19 ± 0.45
TMC (ICLR’21)	74.61 ± 2.64	90.08 ± 0.29
ETMC (TPAMI’22)	75.76 ± 2.83	90.82 ± 0.25
QMF (ICML’23)	77.11 ± 1.46	92.82 ± 0.05
EAU (CVPR’24)	77.30 ± 2.71	92.43 ± 0.12
PDF (ICML’24)	78.26 ± 0.96	93.06 ± 0.24
PML (Ours)	**79.58 ± 0.73**	**93.15 ± 0.09**

Table 3: Accuracy (%) performance on normal test sets. The best and second-best results are in bold and underlined, respectively.

DUA-Nets [Geng *et al.*, 2021], **TMC** [Han *et al.*, 2020], **ETMC** [Han *et al.*, 2022b], **TMDL-OA** [Liu *et al.*, 2022], **DFTMC** [Han *et al.*, 2022a], **DCP-CV** [Lin *et al.*, 2022], **DCP-CG** [Lin *et al.*, 2022], **QMF** [Zhang *et al.*, 2023], **UIMC** [Xie *et al.*, 2023], **PDF** [Cao *et al.*, 2024], **EAU** [Gao *et al.*, 2024], **ECML** [Xu *et al.*, 2024a], **TMNR** [Xu *et al.*, 2024b], and **CCML** [Liu *et al.*, 2024]. As shown in Tables 1 to 3, our method consistently outperforms all baseline methods across all datasets. Non-convergent results for certain methods on specific datasets (*e.g.*, DFTMC on NUSOBJ and Fashion-MV) are marked with ‘-’. More specifically, key observations from the results include:

(1) In the normal testing setting, our PML demonstrates a consistent performance advantage on all datasets, highlighting its superiority and generalizability to various modalities. For example, our method outperforms the best competitor, DCP-CG, by **6.31%** on the LandUse dataset. In addition, on most datasets, our method shows small performance fluctuation, *i.e.*, small std, indicating its promising stability.

(2) From the results of noisy setting, as shown in Table 2, although most competing methods suffer remarkable performance degradation due to noisy and misaligned modalities/views (*e.g.*, QMF and PDF), our approach maintains superior accuracy across all datasets, *e.g.*, NUSOBJ, Scene15, LandUse, and Leaves100. More specifically, on the Leaves100 dataset, our method outperforms the strongest baseline UIMC by **9.44%** far beyond expectations, which is sufficient to demonstrate the effectiveness of vMF-based uncertainty modeling and the robustness of our PML.

4.4 Ablation Study

To further analyze the contribution of each component, we conduct ablation experiments on normal and noisy test sets for key designs of PML. The experimental results are shown in Table 4. From the table, one could see that all components contribute to the performance. On the one hand, with-

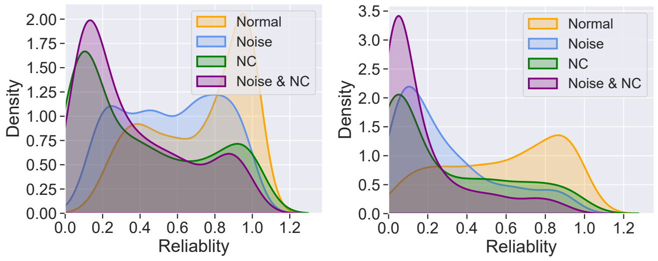
Settings	vMF-PCL	RMF	Handwritten	MSRC-V1	NUSOBJ	Fashion-MV	Scene15	Leaves100
Normal		√	98.85 ± 0.60	96.81 ± 1.65	48.44 ± 0.61	98.26 ± 0.28	82.13 ± 0.84	96.81 ± 1.65
	√		99.30 ± 0.47	99.25 ± 0.56	49.05 ± 0.48	99.10 ± 0.12	82.59 ± 0.60	99.25 ± 0.56
	√	√	**99.32 ± 0.45**	**99.50 ± 0.15**	**49.11 ± 0.39**	**99.12 ± 0.22**	**82.66 ± 0.91**	**99.50 ± 0.15**
Noise		√	92.25 ± 1.43	81.03 ± 3.20	44.78 ± 0.60	88.80 ± 2.08	70.06 ± 1.56	81.03 ± 3.20
	√		98.50 ± 0.68	85.97 ± 2.37	47.01 ± 0.49	94.01 ± 1.38	72.59 ± 1.33	85.97 ± 2.37
	√	√	**98.77 ± 0.48**	**88.47 ± 1.57**	**47.28 ± 0.43**	**96.33 ± 0.21**	**72.71 ± 1.22**	**88.47 ± 1.57**

Table 4: Ablation experimental results on the Handwritten, MSRC-V1, NUSOBJ, Fashion-MV, Scene15, and Leaves100 datasets. The best results are in bold.

out vMF-PCL, the modality-specific information cannot be fully exploited. On the other hand, without RMF, the features of different modalities cannot be effectively and robustly integrated, both leading to inferior performance. Notably, vMF-PCL has a higher contribution to the performance than RMF, indicating that modality-specific information is the basis of fusion to ensure complementary features are extracted from each modality. In summary, the full version of PML achieves the best performance in both normal and noisy settings, which shows that each proposed component plays an important role in the proposed PML.

4.5 Reliability Study

In this section, we present quantitative experiments to visually demonstrate the model's capability to estimate reliability under varying conditions. After training the model on the standard training set, we introduced different types of noise into the test set to assess whether our approach could effectively quantify the associated reliability. The experimental results are illustrated in Figure 3, covering four distinct scenarios: (1) data with normal conditions (*i.e.*, , clean data), (2) data with additive Gaussian noise, (3) data with noisy correspondence (NC), and (4) data subjected to both Gaussian noise and noisy correspondence (Noise & NC). From Equation (2), it can be observed that the introduction of noise leads to a noticeable decline in the model's estimated reliability. Specifically, the model demonstrates high confidence under normal conditions (yellow region), while the reliability estimation deteriorates as noise is added. In the case of Gaussian noise (blue region), the reliability scores exhibit moderate degradation, reflecting our model could correctly capture the uncertainty. Similarly, when noisy correspondence is introduced, the model identifies the inherent uncertainty stemming from incorrect cross-modal relationships, resulting in a more significant reduction in the estimated reliability. The most pronounced impact is observed when both noise sources are present, where the reliability is substantially diminished, indicating that our method successfully captures compounding uncertainties. These findings underscore the effectiveness of our proposed approach in accurately quantifying reliability in the presence of various noise types. The results not only validate the model's ability to differentiate between clean and noisy data but also demonstrate its potential to serve as a trustworthy mechanism for real-world applications where data uncertainty is prevalent.



(a) LandUse

(b) Leaves100

Figure 3: Reliability Density on the test sets with various conditions (normal, Gaussian noise, noisy correspondence, and mixture noise) for the LandUse and Leaves100 datasets.

5 Conclusion

In this paper, we proposed a novel **Probabilistic Multimodal Learning (PML)** framework based on von Mises-Fisher (vMF) distributions to enhance the reliability of multimodal classification. By modeling each data point with a vMF distribution, characterized by a mean direction and a concentration parameter, our method effectively encapsulates discriminative information into the directional representations while capturing the intrinsic data uncertainty. Unlike Gaussian-based models, our PML leverages the concentration parameter to directly measure data uncertainty, providing more stable and robust reliability modeling. To achieve robust multimodal fusion, we propose a Reliable Multimodal Fusion mechanism (RMF) that dynamically balances contributions from different modalities based on the predicted consistency and learned reliability. Extensive experiments on nine benchmark datasets demonstrate the effectiveness and robustness of our PML over 14 state-of-the-art approaches, in both normal and noise settings, showcasing its robustness in handling noisy and misaligned modalities. Furthermore, ablation studies validate the necessity of the key components in our framework, including our vMF-PCL and RMF. While this work focuses on the classic multimodal classification task, in the future, we plan to extend the proposed PML to broader applications such as multimodal clustering, cross-modal retrieval, etc.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2024YFB4710604; in part by NSFC under Grant 62472295, 62176171, U21B2040, 624B2099 and 623B2075; in part by Sichuan Science and Technology Planning Project under Grant 2024NSFTD0047 and 2024NSFTD0038; in part by System of Systems and Artificial Intelligence Laboratory pioneer fund grant; and in part by the Fundamental Research Funds for the Central Universities under Grant CJ202403 and CJ202303.

References

[Afouras *et al.*, 2018] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727, 2018.

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

[Chen *et al.*, 2019] Hao Chen, Youfu Li, and Dan Su. Multimodal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection. *Pattern Recognition*, 86:376–385, 2019.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[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.

[Conti *et al.*, 2022] Jean-Rémy Conti, Nathan Noiry, Stephan Clemencon, Vincent Despiegel, and Stéphane Gentric. Mitigating gender bias in face recognition using the von mises-fisher mixture model. In *International Conference on Machine Learning*, pages 4344–4369. PMLR, 2022.

[Duin, 1998] Robert Duin. Multiple Features. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C5HC70.

[Gao *et al.*, 2024] Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26876–26885, 2024.

[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.

[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.

[Heo *et al.*, 2018] Jay Heo, Hae Beom Lee, Saehoon Kim, Juho Lee, Kwang Joon Kim, Eunho Yang, and Sung Ju Hwang. Uncertainty-aware attention for reliable interpretation and prediction. *Advances in neural information processing systems*, 31, 2018.

[Kendall and Gal, 2017] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Lakshminarayanan *et al.*, 2017] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[Li *et al.*, 2021] Shen Li, Jianqing Xu, Xiaqing Xu, Pengcheng Shen, Shaoxin Li, and Bryan Hooi. Spherical confidence learning for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15629–15637, 2021.

[Lin *et al.*, 2022] Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4447–4461, 2022.

[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] Ying Liu, Lihong Liu, Cai Xu, Xiangyu Song, Ziyu Guan, and Wei Zhao. Dynamic evidence decoupling for trusted multi-view learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7269–7277, 2024.

[Mai *et al.*, 2020] Sijie Mai, Haifeng Hu, and Songlong Xing. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 164–172, 2020.

[Nagrani *et al.*, 2021] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in* *neural information processing systems*, 34:14200–14213, 2021.

[Natarajan *et al.*, 2012] Pradeep Natarajan, Shuang Wu, Shiv Vitaladevuni, Xiaodan Zhuang, Stavros Tsakalidis, Unsang Park, Rohit Prasad, and Premkumar Natarajan. Multimodal feature fusion for robust event detection in web videos. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1298–1305. IEEE, 2012.

[Niu *et al.*, 2016] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. Sentiment analysis on multi-view social data. In *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22*, pages 15–27. Springer, 2016.

[Peng *et al.*, 2022] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247, 2022.

[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.

[Shi and Jain, 2019] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6902–6911, 2019.

[Tang *et al.*, 2022] Xiaolin Tang, Kai Yang, Hong Wang, Jiahang Wu, Yechen Qin, Wenhao Yu, and Dongpu Cao. Prediction-uncertainty-aware decision-making for autonomous vehicles. *IEEE Transactions on Intelligent Vehicles*, 7(4):849–862, 2022.

[Wang *et al.*, 2015] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2015.

[Wen *et al.*, 2023a] Jie Wen, Chengliang Liu, Shijie Deng, Yicheng Liu, Lunke Fei, Ke Yan, and Yong Xu. Deep double incomplete multi-view multi-label learning with incomplete labels and missing views. *IEEE transactions on neural networks and learning systems*, 2023.

[Wen *et al.*, 2023b] Jie Wen, Chengliang Liu, Gehui Xu, Zhihao Wu, Chao Huang, Lunke Fei, and Yong Xu. Highly confident local structure based consensus graph learning for incomplete multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15712–15721, 2023.

[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.

[Xie *et al.*, 2023] Mengyao Xie, Zongbo Han, Changqing Zhang, Yichen Bai, and Qinghua Hu. Exploring and exploiting uncertainty for incomplete multi-view classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19873–19882, 2023.

[Xu *et al.*, 2022] Jie Xu, Yazhou Ren, Huayi Tang, Zhimeng Yang, Lili Pan, Yang Yang, Xiaorong Pu, S Yu Philip, and Lifang He. Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7470–7482, 2022.

[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 Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 5263–5271. International Joint Conferences on Artificial Intelligence Organization, 8 2024. Main Track.

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

[Yang *et al.*, 2022] Mouxing Yang, Yunfan Li, Peng Hu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Robust multi-view clustering with incomplete information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):1055–1069, 2022.

[Yu *et al.*, 2021] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10790–10797, 2021.

[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, 2021] Jiaxin Zhang. Modern monte carlo methods for efficient uncertainty quantification and propagation: A survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(5):e1539, 2021.

[Zou *et al.*, 2024] Ke Zou, Tian Lin, Zongbo Han, Meng Wang, Xuedong Yuan, Haoyu Chen, Changqing Zhang, Xiaojing Shen, and Huazhu Fu. Confidence-aware multi-modality learning for eye disease screening. *Medical Image Analysis*, 96:103214, 2024.