# Provable Dynamic Fusion for Low-Quality Multimodal Data

**Qingyang Zhang** [1] **Haitao Wu** [1] **Changqing Zhang** [1 2] **Qinghua Hu** [1 2]
**Huazhu Fu** [3] **Joey Tianyi Zhou** [3 4] **Xi Peng** [5]

## Abstract

The inherent challenge of multimodal fusion is to precisely capture the cross-modal correlation and flexibly conduct cross-modal interaction. To fully release the value of each modality and mitigate the influence of low-quality multimodal data, dynamic multimodal fusion emerges as a promising learning paradigm. Despite its widespread use, theoretical justifications in this field are still notably lacking. *Can we design a provably robust multimodal fusion method?* This paper provides theoretical understandings to answer this question under a most popular multimodal fusion framework from the generalization perspective. We proceed to reveal that several uncertainty estimation solutions are naturally available to achieve robust multimodal fusion. Then a novel multimodal fusion framework termed Quality-aware Multimodal Fusion (QMF) is proposed, which can improve the performance in terms of classification accuracy and model robustness. Extensive experimental results on multiple benchmarks can support our findings.

## 1. Introduction

Our perception of the world is based on multiple modalities, e.g., touch, sight, hearing, smell and taste. With the development of sensory technology, we can easily collect diverse forms of data for analysis. For example, multi-sensor in autonomous driving and wearable electrical devices (Xiao et al., 2020; Wen et al., 2022), or various examinations in medical diagnosis and treatment (Qiu et al., 2022; Acosta et al., 2022). Intuitively, fusing information from different modalities offers the possibility of exploring cross-modal correlation and gaining better performance. However, conventional fusion methods have largely overlooked the unreliable quality of multimodal data. In real-world settings, the quality of different modalities usually varies due to unexpected environmental issues. Some recent studies have shown both empirically and theoretically that multimodal fusion may fail on low-quality multimodal data, e.g., imbalanced (Wang et al., 2020; Peng et al., 2022; Huang et al., 2022), noisy or even corrupted (Huang et al., 2021b) multimodal data. Empirically, it is recognized that multimodal models cannot always outperform unimodal models especially in a high noise (Scheunders & De Backer, 2007; Eitel et al., 2015; Silva et al., 2022) or imbalanced modality quality (Wu et al., 2022; Peng et al., 2022) regime. Theoretically, the previous study proves that the advantages of multimodal learning may vanish under the setting of limited data volume (Huang et al., 2021a) which implies the exploitation of cross-modal relationship is not a free lunch. To fully release the value of each modality and mitigate the influence of low-quality multimodal data, introducing dynamic fusion mechanism emerges as a promising way to obtain reliable predictions. As a concrete example, previous work (Guan et al., 2019) proposes a dynamic weighting mechanism to depict illumination condition of scenes. By introducing dynamics, they can integrate reliable cues from multi-spectral data for around-the-clock applications (e.g., pedestrian detection in security surveillance and autonomous driving). Dynamic fusion has been used in diverse real-world multimodal applications, including multimodal classification (Han et al., 2021; Geng et al., 2021; Han et al., 2022b), regression (Ma et al., 2021), object detection (Li et al., 2022a; Zhang et al., 2019; Chen et al., 2022b) and semantic segmentation (Tian et al., 2020). While dynamic multimodal fusion shows excellent power in practice, theoretical understanding is notably lack in this field with the following fundamental open problem: *Can we realize reliable multimodal fusion in practice with theoretical guarantee?*

This paper tries to shed light upon the theoretical advantage and criterion of robust multimodal fusion. Following previous works in multimodal learning theory (Huang et al.,

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China [2]Tianjin Key Lab of Machine Learning, Tianjin University, Tianjin, China [3]Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore [4]Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), Singapore [5]College of Computer Science, Sichuan University, Chengdu, China. Correspondence to: Changqing Zhang <zhangchangqing@tju.edu.cn>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

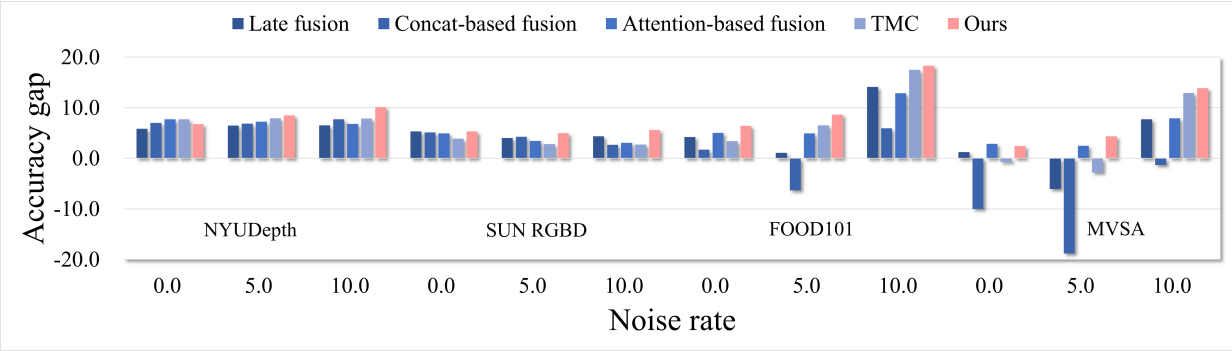



*Figure 1.* Visualization of accuracy gap between multimodal learning methods (e.g., late fusion, align-based fusion, MMTM) and single-modal learning methods using the best modality on noisy multimodal data. Noted that the performance existing multimodal fusion methods degrade significantly of compared with their best unimodal counterparts in a high noise regime, while the proposed QMF consistently outperforms unimodal methods on low-quality data.

2021b; Wang et al., 2020), the framework we study is also abstracted from decision-level multimodal fusion, which is one of the most fundamental research topics in multimodal learning (Baltrušaitis et al., 2018). In particular, we devise a novel Quality-aware Multimodal Fusion (**QMF**) framework for multimodal learning. Key to our framework, we leverage energy-based uncertainty to characterize the quality of each modality. Our contributions can be summarised as follows:

- This paper provides a rigorous theoretical framework to understand the advantage and criterion of robust multimodal fusionas shown in Figure 2. Firstly, we characterize the generalization error bound of decision-level multimodal fusion methods from a Rademacher complexity perspective. Then, we identify under what conditions dynamic fusion outperforms static, i.e., when the fusion weights of multimodal fusion is negatively correlates to the unimodal generalization errors, dynamic fusion methods provably outperform static.
- Under the theoretical analysis, we proceed to reveal that the generalization ability of dynamic fusion coincides with the performance of uncertainty estimation. This directly implies a principle to design and evaluate new dynamic fusion algorithms.
- Directly motivated by the above analysis, we propose a novel dynamic multimodal fusion method termed Quality-aware Multimodal Fusion (**QMF**), which serves as a realization for provably better generalization ability. As shown in Figure 1, extensive experiments on commonly used benchmarks are carried out to empirically validate the theoretical observations.

## 2. Related works

### 2.1. Multimodal Fusion

Multimodal fusion is one of the most original and fundamental topics in multimodal learning, which typically aims to integrate modality-wise features into a joint representation for downstream multimodal learning tasks. Multimodal fusion can be classified into early fusion, intermediate fusion and late fusion. Although studies in neuroscience and machine learning suggest that intermediate fusion could benefit representation learning (Schroeder & Foxe, 2005; Macaluso, 2006), late fusion is still the most widely used method for multimodal learning due to its interpretation and practical simplicity. By introducing modality-level dynamics based on various strategies, dynamic fusion practically improves overall performance. As a concrete example, the previous work (Guan et al., 2019) proposes a dynamic weighting mechanism to depict illumination conditions of scenes. By introducing dynamics, they can integrate reliable cues from multi-spectral data for around-the-clock applications (e.g., pedestrian detection in security surveillance and autonomous driving). Combining with additional dynamic mechanism (e.g., a simple weighting strategy or Dempster-Shafer Evidence Theory (Shafer, 1976)), recent uncertainty-based multimodal fusion methods show remarkable advantages in various tasks, including clustering (Geng et al., 2021), classification (Han et al., 2021; 2022b; Tellamekala et al., 2022; Subedar et al., 2019; Chen et al., 2022a), regression (Ma et al., 2021), object detection (Zhang et al., 2019; Li et al., 2022b) and semantic segmentation (Tian et al., 2020; Chang et al., 2022).

### 2.2. Uncertainty Estimation

Multimodal machine learning has achieved great success in various real-world application. However, the reliability of current fusion methods is still notably unexplored, which limits their application in safety-critic field (e.g., financial risk, medical diagnosis). The motivation of uncertainty estimation is to indicate whether the predictions given by machine learning models are prone to be wrong. Many uncertainty estimation methods have been proposed

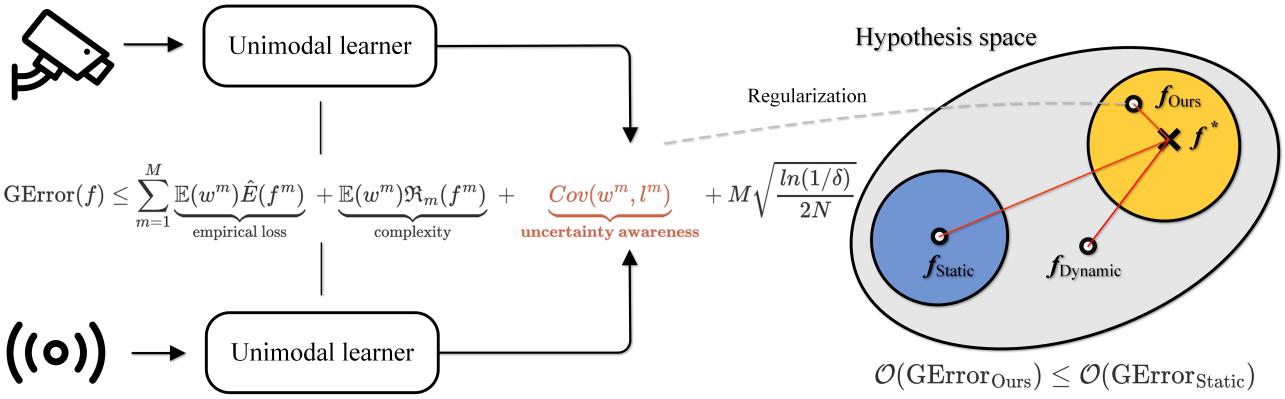



*Figure 2.* **Left:** The generalization error upper bound of multimodal fusion method $f$ can be characterized by its performance on each modality in terms of empirical loss, model complexity and uncertainty awareness. **Right:** Dynamic vs Static multimodal fusion hypothesis space, where the latter is a subset of the former. $f_{\text{static}}$, $f_{\text{dynamic}}$ are the hypothesises of static and dynamic fusion methods respectively and $f^*$ is the true mapping. Informally, closer to the true mapping leads to less error. Under some certain conditions, dynamic multimodal fusion methods (e.g., the proposed QMF) can be well regularized and thus provably achieve better generalization ability.

in the past decades, including Bayesian neural networks (BNNs) (Denker & LeCun, 1990; Mackay, 1992; Neal, 2012) and its varieties (Gal & Ghahramani, 2016; Han et al., 2022a), deep ensembles (Lakshminarayanan et al., 2017; Havasi et al., 2021), predictive confidence (Hendrycks & Gimpel, 2017), Dempster-Shafer thoery (Han et al., 2021) and energy score (Liu et al., 2020). Predictive confidence expects the predicted class probability to be consistent with the empirical accuracy, which is usually referred in classification tasks. Dempster-Shafer theory (DST) is a generalization of Bayesian theory to subjective probabilities and a general framework for modeling epistemic uncertainty. Energy score emerges as a promising way to capture Out-of-Distribution (OOD) uncertainty, which arises when a machine learning model encounters an input that differs from its training data, and thus the output from the model is unreliable. A plethora of recent researches have studied the issue of OOD uncertainty (Ming et al., 2022; Chen et al., 2021; Meinke & Hein, 2019; Hendrycks et al., 2019). In this paper, we investigate predictive confidence, the Dempster-Shafer theory and energy score due to their theoretical interpretability and effectiveness.

## 3. Theory

In this section, we first clarify the basic notations and the formal definition of multimodal fusion used in Section 3.1. Then we provide main theoretical results in Section 3.2 to rigorously demonstrate when and how dynamic fusion methods work from the perspective of generalization ability (Bartlett & Mendelson, 2002). Due to space constraints, we defer the full details to Appendix A and only present a brief summary of the proofs.

### 3.1. Preliminaries

We initialize by introducing the necessary notations for our theoretical frameworks. Considering a learning task on the data $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where $x = \{x^{(1)}, \cdots, x^{(M)}\}$ has $M$ modalities and $y \in \mathcal{Y}$ denotes the data label. The multimodal training data is defined as $D_{\text{train}} = \{x_i, y_i\}_{i=1}^{N}$. Specifically, we use $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ to denote the input space, target space and latent space. Similar to the previous work in multimodal learning theory (Huang et al., 2021a), we define $h : \mathcal{X} \mapsto \mathcal{Z}$ is a multimodal fusion mapping from the input space to the latent space, and $g : \mathcal{Z} \mapsto \mathcal{Y}$ is a task mapping. Our goal is to learn a reliable multimodal model $f = g \circ h(x)$ performing well on the unknown multimodal test dataset $D_{\text{test}}$. $D_{\text{train}}$ and $D_{\text{test}}$ are both drawn from joint distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. Here $f = g \circ h(x)$ represents the composite function of $h$ and $g$.

### 3.2. When and How Dynamic Multimodal Fusion Help

For simplicity, we provide analysis of ensemble-like late fusion strategy using logistic loss function in two-class classification setting. Our analysis follows this roadmap: (1) we first characterize the generalization error bound of dynamic late fusion using Rademacher complexity (Bartlett & Mendelson, 2002) and then separate the bound into three components (Theorem 1); (2) base on above separation, we further prove that dynamic fusion achieves better generalization ability under certain conditions (Theorem 2). We initiate our analysis with the basic setting as follows.

**Basic setting.** Under a $M$ input modalities and two-class classification scenario, we define $f^m$ as the unimodal classifier on modality $x^{(m)}$. The final prediction of late fusion multimodal method is calculated by weighting decisions

from different modalities: $f(x) = \sum_{m=1}^{M} w^m \cdot f^m(x^{(m)})$, where $f(x)$ denotes the final prediction. In contrast to static late fusion, the weights in dynamic multimodal fusion are generated dynamically and vary for different samples. For clarity, we use subscript to distinguish them, i.e., $w^m_{\text{static}}$ refers to the ensemble weight of modality $m$ in static late fusion and $w^m_{\text{dynamic}}$ refers to the weight in dynamic fusion. Specifically, $w^m_{\text{static}}$ is a constant and $w^m_{\text{dynamic}}(\cdot)$ is a function of the input sample $x$. The generalization error of two-class multimodal classifier $f$ is defined as:

$$\text{GError}(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)], \tag{1}$$

where $\mathcal{D}$ is the unknown joint distribution, and $\ell$ is logistic loss function. For convenience, we simplify unimodal classifier loss $\ell(f^m(x^m), y)$ as $l^m$ and omit the inputs in the following analysis. Now we present our first main result regarding multi-modal fusion.

**Theorem 1** (Generalization Bound of Multimodal Fusion)**.** Let $D_{\text{train}} = \{x_i, y_i\}_{i=1}^N$ be a training dataset of $N$ samples, $\hat{E}(f^m)$ is the unimodal empirical errors of $f^m$ on $D_{\text{train}}$. Then for any hypothesis $f$ in $\mathcal{H}$ (i.e., $\mathcal{H} : \mathcal{X} \to \{-1, 1\}$, $f \in \mathcal{H}$) and $1 > \delta > 0$, with probability at least $1 - \delta$, it holds that

$$\text{GError}(f) \leq \underbrace{\sum_{m=1}^{M} \mathbb{E}(w^m)\hat{E}(f^m)}_{\text{Term-L (average empirical loss)}} + \underbrace{\sum_{m=1}^{M} \mathbb{E}(w^m)\mathfrak{R}_m(f^m)}_{\text{Term-C (average complexity)}} + \underbrace{\sum_{m=1}^{M} Cov(w^m, l^m)}_{\text{Term-Cov (covariance)}} + M\sqrt{\frac{ln(1/\delta)}{2N}}, \tag{2}$$

where $\mathbb{E}(w^m)$ is the expectations of fusion weights on joint distribution $\mathcal{D}$, $\mathfrak{R}_m(f^m)$ is Rademacher complexity, $Cov(w^m, \ell^m)$ is the covariance between fusion weight and loss.

Intuitively, Theorem 1 demonstrates that the generalization error of multimodal classifier is bounded by the weighted average performances of all the unimodal classifiers in terms of empirical loss, model complexity and the covariance between fusion weight and unimodal loss. Having established the general error bound, our next goal is to verify when dynamic multimodal late fusion indeed achieves tighter bound than that of static late fusion. Informally, in Eq. 1, Term-Cov measures the joint variability of $w^m$ and $\ell^m$. Remember that in static multimodal fusion $w^m_{\text{static}}$ is a constant, which means Term-Cov $= 0$ for any static fusion methods. Thus the generalization error bound of static fusion methods reduces to

$$\text{GError}(f_{\text{static}}) \leq \underbrace{\sum_{m=1}^{M} w^m_{\text{static}}\hat{E}(f^m)}_{\text{Term-L (average empirical loss)}} + \underbrace{\sum_{m=1}^{M} w^m_{\text{static}}\mathfrak{R}_m(f^m)}_{\text{Term-C (average complexity)}} + M\sqrt{\frac{ln(1/\delta)}{2N}}. \tag{3}$$

So when summation of Term-L , Term-C is invariant or smaller in dynamic fusion and Term-Cov $\leq 0$, we can ensure that dynamic fusion provably outperforms static fusion. This theorem is formally presented as

**Theorem 2.** Let $\mathcal{O}(\text{GError}(f_{\text{dynamic}}))$, $\mathcal{O}(\text{GError}(f_{\text{static}}))$ be the upper bound of generalization error of multimodal classifier using dynamic and static fusion strategy respectively. $\hat{E}(f^m)$ is the unimodal empirical errors of $f^m$ on $D_{\text{train}}$ defined in Theorem 1. Then for any hypothesis $f_{\text{dynamic}}, f_{\text{static}}$ in $\mathcal{H} : \mathcal{X} \to \{-1, 1\}$ and $1 > \delta > 0$, it holds that

$$\mathcal{O}(\text{GError}(f_{\text{dynamic}})) \leq \mathcal{O}(\text{GError}(f_{\text{static}})) \tag{4}$$

with probability at least $1 - \delta$, if we have

$$\mathbb{E}(w^m_{\text{dynamic}}) = w^m_{\text{static}} \tag{5}$$

and

$$r(w^m_{\text{dynamic}}, \ell(f^m)) \leq 0 \tag{6}$$

for all input modalities, where $r$ is the Pearson correlation coefficient which measures the correlation between fusion weights $w^m_{\text{dynamic}}$ and unimodal loss $\ell^m$.

**Remark.** Theoretically, optimizing over the same function class efficiently results in the same empirical loss. Suppose for each modality $m$, the unimodal classifier $f^m$ we used in dynamic and static fusion are of the same architecture, then the intrinsic complexity of unimodal classifier $\mathfrak{R}_m(f^m)$ and empirical risk $\hat{E}(f^m)$ can be invariant. Thus in this case, it holds that

$$\sum_{m=1}^{M} \mathbb{E}(w^m_{\text{dynamic}})\hat{E}(f^m) \leq \sum_{m=1}^{M} w^m_{\text{static}}\hat{E}(f^m), \tag{7}$$

and

$$\sum_{m=1}^{M} \mathbb{E}(w^m_{\text{dynamic}})\mathfrak{R}_m(f^m) \leq \sum_{m=1}^{M} w^m_{\text{static}}\mathfrak{R}_m(f^m), \tag{8}$$

if Eq. 5 is satisfied for any modality $m$. According to Theorem 2, it is easy to derive the conclusion that the main

challenge of achieving reliable dynamic multimodal fusion is to learn a reasonable $w^m_{\text{dynamic}}(x)$ for each modality that satisfies Eq. 5 and Eq. 6.

## 4. Method

Now we proceed to answer "How to realize robust dynamic fusion?". In this section, we theoretically identify the connection between dynamic multimodal fusion and uncertainty estimation. Then, a unified dynamic multimodal fusion framework termed Quality-aware Multimodal Fusion (QMF) is proposed. We next show how to realize this framework in decision-level late fusion and classification tasks to support our findings.

### 4.1. Coincidence with Uncertainty Estimation

Firstly, we focus on how to satisfy Eq. 6. As we discuss in Section 2.2, the common motivation of various uncertainty estimation methods is to provide an indicator of whether the predictions given by models are prone to be wrong. This motivation is inherently close to obtaining weights that satisfy Eq. 6. We formulate this claim with the following assumption

**Assumption 1.** *Given an effective uncertainty estimator $u^m : \mathcal{X} \to \mathbb{R}$ on modality $m$, the estimated uncertainty $u^m(x)$ is positively correlated with its modal-specific loss $\ell^m(x)$: $r(u^m, \ell^m(x)) \geq 0$, where $r$ is the Pearson correlation coefficient.*

This insight offers opportunity to explore novel dynamic fusion methods provably outperform conventional static fusion methods. Similar to previous dynamic fusion methods (Blundell et al., 2015; Zhang et al., 2019; Han et al., 2022b), we deploy modal-level weighting strategy to introduce dynamics.

**Uncertainty-aware weighting.** The uncertainty-aware fusion weighting $w^m : \mathcal{X} \to \mathbb{R}$ is a function that linearly and negatively relates to the corresponding uncertainty

$$w^m(x) = \alpha^m \, u^m(x) + \beta^m, \tag{9}$$

where $\alpha^m < 0$, $\beta^m \geq 0$ are modal-specific hyper-parameters. $u^m(x)$ is the uncertainty of modality $m$. By tuning hyper-parameters $\alpha^m$, $\beta^m$, we can ensure dynamic fusion weights satisfied Eq. 5 and 6 simultaneously. This lemma is formally presented as

**Lemma 1 (Satisfiability).** With Assumption 1, for any $w^m_{\text{static}} \in \mathbb{R}$, there always exist $\beta^m \in \mathbb{R}$ such that

$$\mathbb{E}(w^m_{\text{dynamic}}) = w^m_{\text{static}}, r(w^m_{\text{dynamic}}, \ell(f^m)) \leq 0. \tag{10}$$

Once we obtain the fusion weights, we can perform uncertainty-aware weighting fusion in decision-level according to the following rule

$$f(x) = \sum_{m=1}^{M} w^m(x) \cdot f^m(x), \tag{11}$$

where $f^m(x)$ defined in Section 3.2 denotes unimodal prediction on modality $m$.

### 4.2. Enhance Correlation by Additional Regularization

With the above analysis, the core challenges of robust dynamic multimodal fusion present in Section. 3.2 have been reduced to obtain an effective uncertainty estimator in Assumption 1. In our implementation, we leverage energy score (Liu et al., 2020), which is a widely accepted metric in the literature of uncertainty learning. Energy score [1] bridges the gap between the Helmholtz free energy of a given data point and its density. For multimodal data, the density functions of different modalities can be estimated by the corresponding energy function:

$$log\ p(x^{(m)}) = -\text{Energy}(x; f^m)/\mathcal{T}^m - log\ Z^m, \tag{12}$$

where $x^{(m)}$ is the $m$-th input modality and $f^m$ is the unimodal classification model. Energy$(\cdot)$ is the energy function and $Z^m$ is an intractable constant for all $x^m$. The above equation suggests that $-$Energy$(x^{(m)}; f^m)$ is linearly aligned with density $p(x^{(m)})$. The energy score for the $m$-th modality of input $x$ can be calculated as

$$\text{Energy}(x^{(m)}) = -\mathcal{T}^m \cdot log \sum_{k}^{K} e^{f^m_k(x^{(m)})/\mathcal{T}^m}, \tag{13}$$

where $f^m_k(x^{(m)})$ is the output logits of classifier $f^m$ corresponding to the $k$-th class label and $\mathcal{T}^m$ is a temperature parameter. Intuitively, more uniformly distributed prediction leads to higher estimated uncertainty.

However, it has been shown experimentally that the uncertainty estimated in this way without additional regularization is not well enough to satisfy our Assumption 1. To address this, we propose a sampling-based regularization technology to enhance the original method in terms of correlation. The most simple and straightforward way to improve the correlation between estimated uncertainty and respective loss is to leverage the sample-wise loss during training stage as supervision information. However, due to the over-parameterization phenomenon of deep neural networks, the

[1] While another line of previous works usually incorporate an auxiliary outlier dataset (e.g., random noised out-of-distribution data) during training for higher performance, for clarity and a strictly fair comparison, we conduct our experiments without the help of additional data.

**Algorithm 1** Training Pseudo Code of Quality-aware Multimodal Fusion (QMF)

**Input :** Multimodal training dataset $D_{\text{train}}$, the number of sampling $\mathrm{T}$, hyperparameters $\lambda$, temperature parameters $\{\mathcal{T}^m\}_{m=1}^M$, unimodal predictors $\{f^m(\cdot)\}_{i=m}^M$;

**Output :** The multimodal classifier $f$;

**for** *each iteration* **do**

  Obtain training sample $(x_i, y_i)$ from dataset $D_{\text{train}}$ and the decisions on each modality $f^m(x)$;

  Calculate uncertainty-aware fusion weights $[w_i^1, \cdots, w_i^m]$ defined in Eq. 9;

  Update the average training loss $\kappa_i^m$ of each modalities;

  Obtain the multimodal decision by weighting unimodal predictions dynamically according to Eq. 11;

  Update model parameters of each unimodal predictor by minimizing $\mathcal{L}_{\text{overall}}$ in Eq. 18.

**end**

losses constantly reduce to zero during training. Inspired by recent works in Bayesian learning (Maddox et al., 2019) and uncertainty estimation (Moon et al., 2020; Han et al., 2022a), we propose to leverage the information from historical training trajectory to regularize the fusion weights. Specifically, given the $m$-th modality of a sample $(x_i, y_i)$, the training average loss for $x_i^m$ is calculated as:

$$\kappa_i^m = \frac{1}{\mathrm{T}} \sum_{\mathrm{t}=\mathrm{T_s}}^{\mathrm{T_s}+\mathrm{T}} \ell(y_i, f_{\theta_t}^m(x_i)), \tag{14}$$

where $f_{\theta_t}^m$ is the unimodal classifier on each iteration epoch $t$ with parameters $\theta_t$. After training $\mathrm{T_s} - 1$ epochs, we sample $\mathrm{T}$ times and calculate the average training loss.

Empirically, recent works (Geifman et al., 2019) shown that easy-to-classify samples are learned earlier during training compared to hard-to-classify samples (e.g., noise samples (Arazo et al., 2019)). It is desirable to regularize a dynamic fusion model by learning the following relationship during training

$$\kappa_i^m \geq \kappa_j^m \iff w_i^m \leq w_j^m. \tag{15}$$

We now present the full definition of our regularization term as follows

$$\mathcal{L}_{\text{reg}} = max(0, g(w_i^m, w_j^m)(\kappa_i^m - \kappa_j^m) + |w_i^m - w_j^m|), \tag{16}$$

where

$$g(w_i^m, w_j^m) = \begin{cases} 1 \text{ if } w_i^m > w_j^m, \\ 0 \text{ if } w_i^m = w_j^m, \\ -1 \text{ otherwise.} \end{cases} \tag{17}$$

Inspired by multi-task learning, we define the total loss function as a summation of standard cross-entropy classification losses of multiple modalities and the regularization term

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{CE}}(y, f(x)) + \sum_{m=1}^{M} \mathcal{L}_{\text{CE}}(y, f^m(x^m)) + \lambda \mathcal{L}_{\text{reg}}, \tag{18}$$

where $\lambda$ is a hyperparamter which controls the strength of regularization, $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{reg}}$ are the cross-entropy loss and reguralization term respectively. The whole training process is shown in Algorithm 1.

**Intuitive explanation of the effectiveness of QMF.** Without loss of generality, we assume modality $x^A$ is clean and modality $x^B$ is noisy due to unknown environmental factors or sensor failure. At this time, $x^A$ is in the distribution of clean training data but $x^B$ deviates significantly from it. Accordingly, we have $u(x^A) \leq u(x^B)$ and thus $w^A \geq w^B$. Therefore, for our QMF, the multimodal decision will tend to more rely on the high-quality modality $x^A$ than the other modality $x^B$. By dynamically determining the fusion weights of each modality, the influence of the unreliable modalities can be alleviated.

## 5. Experiment

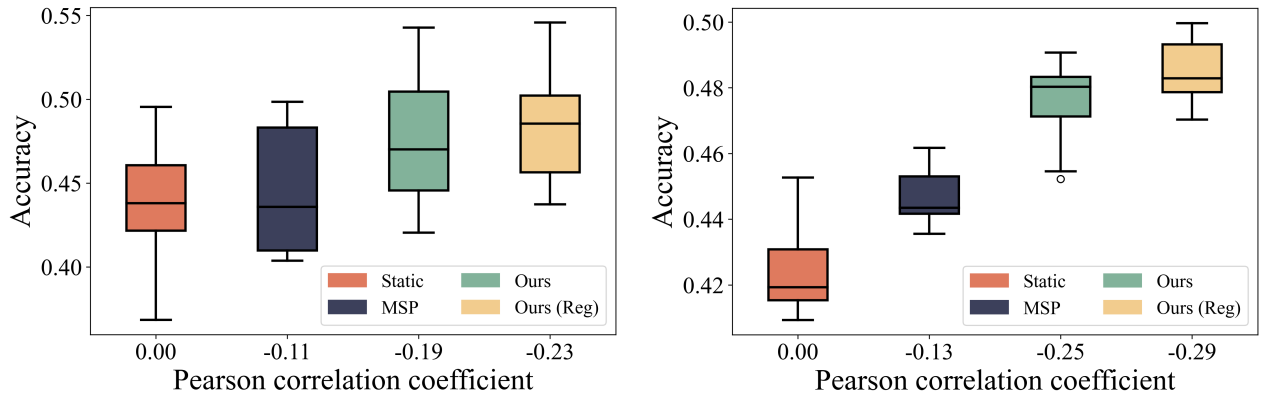
In this section, we conduct experiments on multiple datasets of diverse applications [2]. The main questions to be verified are highlighted here:

- Q1 Effectiveness I. Does the proposed method has better generalization ability than its counterparts? (Support Theorem 1)
- Q2 Effectiveness II. Under what conditions does uncertainty-aware dynamic multimodal fusion work? (Support Theorem 2)
- Q3 Reliability. Does the proposed method have an effective perception for the uncertainty of modality? (Support Assumption 1)
- Q4 Ablation study. What is the key factor of performance improvement in our method?

### 5.1. Experimental Setup

We briefly present the experimental setup here, including the experimental datasets and comparison methods. Please

[2] Code is available at https://github.com/QingyangZhang/QMF.




(a) NYU Depth V2

(b) SUN RGB-D

*Figure 3.* Test accuracy and Pearson correlation coefficient achieved by different fusion methods over 10 times random experiments. The average and worst-case accuracy are highly consistency with uncertainty estimation ability.

refer to Appendix B for more detailed setup.

**Tasks and datasets.** We evaluate our method on two multimodal classification tasks. ◦ Scenes Recognition: NYU Depth V2 (Silberman et al., 2012) and SUN RGB-D (Song et al., 2015) are two public indoor scenes recognition datasets, which are associated with two modalities, i.e., RGB and depth images. ◦ Image-text classification: The UPMC FOOD101 dataset (Wang et al., 2015) contains (possibly noisy) images obtained by Google Image Search and corresponding textual descriptions. MVSA sentiment analysis dataset (Niu et al., 2016) includes a set of image-text pairs with manual annotations collected from social media. Although the datasets above are all under the condition that $M = 2$, it is intuitive and easy to generalize to $M \geq 3$.

**Evaluation metrics.** Due to the randomness involved, we report the mean accuracy, standard deviation and worst-case accuracy on NYU Depth V2 and SUN RGB-D over 10 different seeds. To be consistent with existing works (Han et al., 2022c; Kiela et al., 2019; Yadav & Vishwakarma, 2023), we repeat experiments over 3 times on UMPC FOOD101 and 5 times on MVSA.

**Compared methods.** For scene recognition task, we compare the proposed method with three static fusion methods: Late fusion, Concatenate-based fusion, Alignment-based fusion methods (Wang et al., 2016) and two representative dynamic fusion methods, i.e., MMTM (Joze et al., 2020) and TMC[3] (Han et al., 2021). For image-text classification, we compare against strong unimodal baselines (i.e., Bow, Bert and ResNet-152) as well as sophisticated multimodal fusion methods, including Late fusion, ConcatBow, ConcatBERT and recent sota MMBT (Kiela et al., 2019).

[3]There are two variants in (Han et al., 2021): TMC and ETMC (with additional concatenated-based multimodal fusion strategy). TMC has comparable performance and is a more fair comparison.

### 5.2. Experimental Results

**Classification robustness (Q1).** To validate the robustness of the uncertainty-aware weighting fusion, we evaluate QMF and the compared methods in terms of average and worst-case accuracy under Gaussian noise (for image modality) and blank noise (for text modality) following previous works (Han et al., 2021; Ma et al., 2021; Verma et al., 2021; Hu et al., 2019; Xie et al., 2017). More results under different types of noise (e.g. Salt-Pepper Noise) can be found in Appendix C.2. The experimental results are presented in Table 1. It is observed that QMF usually performs in the top three in terms of both average and worst-case accuracy. This observation indicates that QMF has better generalization ability than their counterparts experimentally. It is also noteworthy that the QMF outperforms the prior **state-of-the-art** methods (i.e., MMBT and TMC) on large-scale benchmark UPMC FOOD101, which illustrates the superiority of the proposed method.

**Connection to uncertainty estimation (Q2).** We further conduct comparisons with QMF realized by various uncertainty estimation algorithms, i.e., prediction confidence (Hendrycks & Gimpel, 2017) and Dempster-Shafer evidence theory (DST) (Han et al., 2021). According to comparison results shown in the Figure 3, it is clear that (i) the generalization ability (i.e., average and worst-case accuracy) of dynamic fusion methods coincide with their uncertainty estimation ability and (ii) our QMF achieves the best performance in terms of classification accuracy and uncertainty estimation in the meantime. This comparison reveals the underlying reason of why QMF outperforms other fusion methods and supports Theorem 2. We show the results on NYU Depth V2 and SUN RGB-D under Gaussian

*Table 1.* Classification comparison when 50% of the modalities are corrupted with Gaussian noise i.e., zero mean with variance of $\epsilon$. The best three results are in bold brown and the best results are highlighted in bold blue. Full results with standard deviation are in Appendix.

| Dataset | Dynamic | Method | $\epsilon = 0.0$ | | $\epsilon = 5.0$ | | $\epsilon = 10.0$ | |
|---|---|---|---|---|---|---|---|---|
| | | | Avg. | Worst. | Avg. | Worst. | Avg. | Worst. |
| NYU Depth V2 | ✗ | RGB | 63.30 | 62.54 | 53.12 | 50.31 | 45.46 | 42.20 |
| | ✗ | Depth | 62.65 | 61.01 | 50.95 | 42.81 | 44.13 | 35.93 |
| | ✗ | Late fusion | 69.14 | 68.35 | 59.63 | 53.98 | 51.99 | 44.95 |
| | ✗ | Concat | 70.30 | **69.42** | 59.97 | 55.89 | **53.20** | **47.71** |
| | ✗ | Align | **70.31** | 68.50 | 59.47 | 56.27 | 51.74 | 44.19 |
| | ✓ | MMTM | **71.04** | **70.18** | **60.37** | **56.73** | 52.28 | 46.18 |
| | ✓ | TMC | **71.06** | **69.57** | **61.04** | **58.72** | **53.36** | **49.23** |
| | ✓ | Ours | 70.09 | 68.81 | **61.62** | **58.87** | **55.60** | **51.07** |
| SUN RGB-D | ✗ | RGB | 56.78 | 56.51 | 48.40 | 47.16 | 42.94 | 41.02 |
| | ✗ | Depth | 52.99 | 51.32 | 37.81 | 35.63 | 33.07 | 30.41 |
| | ✗ | Late fusion | **62.09** | 60.55 | **52.44** | **50.83** | **47.33** | **44.60** |
| | ✗ | Concat | **61.90** | **61.19** | **52.69** | 50.61 | 45.64 | 42.95 |
| | ✗ | Align | 61.12 | 60.12 | 50.05 | 47.63 | 44.19 | 38.12 |
| | ✓ | MMTM | 61.72 | **60.94** | 51.86 | **50.80** | **46.03** | **44.28** |
| | ✓ | TMC | 60.68 | 60.31 | 51.24 | 49.45 | 45.66 | 41.60 |
| | ✓ | Ours | **62.09** | **61.30** | **53.40** | **52.07** | **48.58** | **47.50** |
| FOOD 101 | ✗ | Bow | 82.50 | 82.32 | 61.68 | 60.98 | 41.95 | 41.41 |
| | ✗ | Img | 64.62 | 64.22 | 34.72 | 34.19 | 33.03 | 32.67 |
| | ✗ | Bert | 86.46 | 86.42 | 67.38 | 67.19 | 43.88 | 43.56 |
| | ✗ | Late fusion | **90.69** | **90.58** | 68.49 | 65.05 | **58.00** | 55.77 |
| | ✗ | ConcatBow | 70.77 | 70.68 | 38.28 | 37.95 | 35.68 | 34.92 |
| | ✗ | ConcatBert | 88.20 | 87.81 | 61.10 | 59.25 | 49.86 | 47.79 |
| | ✓ | MMBT | **91.52** | **91.38** | **72.32** | **71.78** | 56.75 | **56.21** |
| | ✓ | TMC | 89.86 | 89.80 | **73.93** | **73.64** | **61.37** | **61.10** |
| | ✓ | Ours | **92.92** | **92.72** | **76.03** | **74.68** | **62.21** | **61.76** |
| MVSA | ✗ | Bow | 48.79 | 35.45 | 42.20 | 32.56 | 41.57 | 32.18 |
| | ✗ | Img | 64.12 | 62.04 | 49.36 | 45.67 | 45.00 | 39.31 |
| | ✗ | Bert | 75.61 | **74.76** | **69.50** | **65.70** | 47.41 | 45.86 |
| | ✗ | Late fusion | **76.88** | 74.76 | 63.46 | 58.57 | 55.16 | 47.78 |
| | ✗ | ConcatBow | 64.09 | 62.04 | 49.95 | 45.28 | 45.40 | 40.95 |
| | ✗ | ConcatBert | 65.59 | 64.74 | 50.70 | 44.70 | 46.12 | 41.81 |
| | ✓ | MMBT | **78.50** | **78.04** | **71.99** | **69.94** | **55.35** | **52.22** |
| | ✓ | TMC | 74.88 | 71.10 | 66.72 | **60.12** | 60.36 | **53.37** |
| | ✓ | Ours | **78.07** | **76.30** | **73.85** | **71.10** | **61.28** | **57.61** |

*Table 2.* Ablation study on NYU Depth V2. Full results with standard deviation are in Appendix C.1.

| UAW | $\mathcal{L}_{\text{reg}}$ | $\epsilon = 0.0$ | | $\epsilon = 5.0$ | | $\epsilon = 10.0$ | | $\epsilon = 20.0$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | Avg. | Worst. | Avg. | Worst. | Avg. | Worst. | Avg. | Worst. |
| ✗ | ✗ | 69.14 | 68.35 | 59.62 | 53.98 | 51.94 | 44.95 | 43.76 | 36.85 |
| ✗ | ✓ | 69.68 | 67.74 | 61.35 | 58.26 | 55.44 | **51.53** | 47.32 | 42.97 |
| ✓ | ✗ | 70.06 | **69.11** | 61.59 | 57.49 | 55.14 | 50.15 | 47.46 | 42.05 |
| ✓ | ✓ | **70.09** | 68.81 | **61.62** | **58.87** | **55.81** | 51.07 | **48.26** | **43.73** |

*Table 3.* Pearson correlation coefficient $r$ between losses and fusion weights of test samples (a higher $|r|$ indicates a better uncertainty estimation).

| | $\epsilon = 0.0$ | $\epsilon = 5.0$ | $\epsilon = 10.0$ |
|---|---|---|---|
| MSP | 0.391 | 0.433 | 0.486 |
| Energy score | 0.272 | 0.429 | 0.510 |
| Entropy | 0.397 | 0.420 | 0.452 |
| Evidence | 0.157 | 0.136 | 0.265 |
| Ours | **0.498** | **0.652** | **0.735** |

noise with zero mean and variance of 10.

**Reliability of QMF (Q3).** We calculate the fusion weights defined in Eq. 9 of different modalities in Table 3 on UPMC FOOD-101. It is observed that the fusion weights of QMF have the most effective perception of modal quality compared with other uncertainty estimation methods (in terms of correlation). This observation justifies our expectation of uncertainty-aware weights in Eq. 9.

**Ablation study (Q4).** We compare different combinations of components (i.e., uncertainty-aware weighting and the regularization term $\mathcal{L}_{\text{reg}}$). Here we also employ Gaussian noise on NYU Depth V2 in Table 2, and more results can be found in the Appendix C.1. It is easy to conclude that 1) adding $\mathcal{L}_{\text{reg}}$ is beneficial to obtaining more reasonable fusion weights; 2) the best performance could be expected with the full QMF. Please refer to Table. 4 in Appendix C.1 for full results with standard deviation.

In summary, the empirical results can support our theoretical findings. These works identify the causes and conditions of performance gains of dynamic multimodal fusion methods. The proposed method can help to improve robustness on multiple datasets.

## 6. Limitations

Even though the proposed method achieves superior performance, there are still some potential limitations. Firstly, the fusion weights of QMF are based on uncertainty estimation, which can be a challenging task in the real world. For example, in our experiments, we can only achieve mild Pearson's $r$ on NYU Depth V2 and SUN RGB-D dataset. Therefore, it is important and valuable to explore novel uncertainty estimation methods in the future work. Secondly, though we characterize the generalization error bound of the proposed method, our theoretical justifications are based on Assumption 1. However, previous work (Fang et al., 2022) reveals that OOD detection is not learnable under some scenarios. Thus it's still a challenging open problem to further characterize the generalization ability of dynamic multimodal fusion.

## 7. Conclusions and Future works

Introducing dynamics in multimodal fusion has yielded remarkable empirical results in various applications, including image classification, object detection and semantic segmentation. Many state-of-the-art multimodal models introduce dynamic fusion strategies, but the inductive bias provided by this technique is not well understood. In this paper, we provide rigorous analysis towards understanding when and what dynamic multimodal fusion methods are more robust on multimodal data in the wild. These findings demonstrate the connection between uncertainty learning and robust multimodal fusion, which further implies a principle to design novel dynamic multimodal fusion methods. Finally, we perform extensive experiments on multiple benchmarks to support our findings. In the work, the energy-based weighting strategy is devised, and other uncertainty estimation ways could be explored. Another interesting direction is proving the dynamic fusion under a more general setting.

## Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (Grant No. 61976151) and A*STAR Central Research Fund. We gratefully acknowledge the support of MindSpore and CAAI. The authors would like to thank Zhipeng Liang (Hong Kong University of Science and Technology) for checking on math details and Zongbo Han, Huan Ma (Tianjin University) for their comments on writing. The authors also appreciate the suggestions from ICML anonymous peer reviewers.

## References


Acosta, J. N., Falcone, G. J., Rajpurkar, P., and Topol, E. J. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784, 2022.

Arazo, E., Ortego, D., Albert, P., O'Connor, N., and McGuinness, K. Unsupervised label noise modeling and loss correction. In *International Conference on Machine Learning*, pp. 312–321. PMLR, 2019.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41 (2):423–443, 2018.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 2002.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural network. In *International Conference on Machine Learning*, 2015.

Chang, Y., Xue, F., Sheng, F., Liang, W., and Ming, A. Fast road segmentation via uncertainty-aware symmetric network. In *International Conference on Robotics and Automation*, 2022.

Chen, J., Li, Y., Wu, X., Liang, Y., and Jha, S. Atom: Robustifying out-of-distribution detection using outlier mining. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2021.

Chen, S., Yang, X., Chen, Y., Yu, H., and Cai, H. Uncertainty-based fusion netwok for automatic skin lesion diagnosis. In *International Conference on Bioinformatics and Biomedicine*, 2022a.

Chen, Y.-T., Shi, J., Ye, Z., Mertz, C., Ramanan, D., and Kong, S. Multimodal object detection via probabilistic ensembling. In *European Conference on Computer Vision*, 2022b.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2009.

Denker, J. and LeCun, Y. Transforming neural-net output levels to probability distributions. In *Advances in Neural Information Processing Systems*, 1990.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Eitel, A., Springenberg, J. T., Spinello, L., Riedmiller, M., and Burgard, W. Multimodal deep learning for robust rgb-d object recognition. In *International Conference on Intelligent Robots and Systems*, 2015.

Fang, Z., Li, Y., Lu, J., Dong, J., Han, B., and Liu, F. Is out-of-distribution detection learnable? In *Advances in Neural Information Processing Systems*, 2022.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2016.

Geifman, Y., Uziel, G., and El-Yaniv, R. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations*, 2019.

Geng, Y., Han, Z., Zhang, C., and Hu, Q. Uncertainty-aware multi-view representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Guan, D., Cao, Y., Yang, J., Cao, Y., and Yang, M. Y. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection. *Information Fusion*, 50:148–157, 2019.

Han, Z., Zhang, C., Fu, H., and Zhou, J. T. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021.

Han, Z., Liang, Z., Yang, F., Liu, L., Li, L., Bian, Y., Zhao, P., Wu, B., Zhang, C., and Yao, J. Umix: Improving importance weighting for subpopulation shift via uncertainty-aware mixup. In *Advances in Neural Information Processing Systems*, 2022a.

Han, Z., Yang, F., Huang, J., Zhang, C., and Yao, J. Multimodal dynamics: Dynamical fusion for trustworthy multimodal classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022b.

Han, Z., Zhang, C., Fu, H., and Zhou, J. T. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 2022c.

Havasi, M., Jenatton, R., Fort, S., Liu, J. Z., Snoek, J., Lakshminarayanan, B., Dai, A. M., and Tran, D. Training independent subnetworks for robust prediction. In *International Conference on Learning Representations*, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

Hu, Z., Tan, B., Salakhutdinov, R. R., Mitchell, T. M., and Xing, E. P. Learning data manipulation for augmentation and weighting. In *Advances in Neural Information Processing Systems*, 2019.

Huang, Y., Du, C., Xue, Z., Chen, X., Zhao, H., and Huang, L. What makes multi-modal learning better than single (provably). In *Advances in Neural Information Processing Systems*, 2021a.

Huang, Y., Lin, J., Zhou, C., Yang, H., and Huang, L. Modality competition: What makes joint training of multimodal network fail in deep learning?(provably). In *International Conference on Machine Learning*, 2022.

Huang, Z., Niu, G., Liu, X., Ding, W., Xiao, X., Wu, H., and Peng, X. Learning with noisy correspondence for cross-modal matching. In *Advances in Neural Information Processing Systems*, 2021b.

Joze, H. R. V., Shaban, A., Iuzzolino, M. L., and Koishida, K. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Kiela, D., Bhooshan, S., Firooz, H., Perez, E., and Testuggine, D. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Li, B., Han, Z., Li, H., Fu, H., and Zhang, C. Trustworthy long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022a.

Li, Q., Zhang, C., Hu, Q., Fu, H., and Zhu, P. Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection. *IEEE Transactions on Multimedia*, 2022b.

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020.

Ma, H., Han, Z., Zhang, C., Fu, H., Zhou, J. T., and Hu, Q. Trustworthy multimodal regression with mixture of normal-inverse gamma distributions. In *Advances in Neural Information Processing Systems*, 2021.

Macaluso, E. Multisensory processing in sensory-specific cortical areas. *The neuroscientist*, 12(4):327–338, 2006.

Mackay, D. J. C. *Bayesian methods for adaptive models*. California Institute of Technology, 1992.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. A simple baseline for bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems*, 2019.

Meinke, A. and Hein, M. Towards neural networks that provably know when they don't know. *arXiv preprint arXiv:1909.12180*, 2019.

Ming, Y., Fan, Y., and Li, Y. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning*, 2022.

Moon, J., Kim, J., Shin, Y., and Hwang, S. Confidence-aware learning for deep neural networks. In *International Conference on Machine Learning*, 2020.

Neal, R. M. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Niu, T., Zhu, S., Pang, L., and El Saddik, A. 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*, pp. 15–27. Springer, 2016.

Peng, X., Wei, Y., Deng, A., Wang, D., and Hu, D. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, 2014.

Qiu, S., Miller, M. I., Joshi, P. S., Lee, J. C., Xue, C., Ni, Y., Wang, Y., Anda-Duran, D., Hwang, P. H., Cramer, J. A., et al. Multimodal deep learning for alzheimer's disease dementia assessment. *Nature communications*, 13(1):1–17, 2022.

Scheunders, P. and De Backer, S. Wavelet denoising of multicomponent images using gaussian scale mixture models and a noise-free image as priors. *IEEE Transactions on Image Processing*, 2007.

Schroeder, C. E. and Foxe, J. Multisensory contributions to low-level,'unisensory'processing. *Current opinion in neurobiology*, 15(4):454–458, 2005.

Shafer, G. *A mathematical theory of evidence*, volume 42. Princeton university press, 1976.

Silberman, N., Hoiem, D., Kohli, P., and Fergus, R. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, 2012.

Silva, A., Luo, L., Karunasekera, S., and Leckie, C. Noise-robust learning from multiple unsupervised sources of inferred labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Song, S., Lichtenberg, S. P., and Xiao, J. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Subedar, M., Krishnan, R., Meyer, P. L., Tickoo, O., and Huang, J. Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

Tellamekala, M. K., Amiriparian, S., Schuller, B. W., André, E., Giesbrecht, T., and Valstar, M. Cold fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. *arXiv preprint arXiv:2206.05833*, 2022.

Tian, J., Cheung, W., Glaser, N., Liu, Y.-C., and Kira, Z. Uno: Uncertainty-aware noisy-or multimodal fusion for unanticipated input degradation. In *International Conference on Robotics and Automation*, 2020.

Verma, V., Qu, M., Kawaguchi, K., Lamb, A., Bengio, Y., Kannala, J., and Tang, J. Graphmix: Improved training of gnns for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2021.

Wang, J., Wang, Z., Tao, D., See, S., and Wang, G. Learning common and specific features for rgb-d semantic segmentation with deconvolutional networks. In *European Conference on Computer Vision*, 2016.

Wang, W., Tran, D., and Feiszli, M. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Wang, X., Kumar, D., Thome, N., Cord, M., and Precioso, F. Recipe recognition with large multimodal food dataset. In *International Conference on Multimedia & Expo Workshops*. IEEE, 2015.

Wen, L., Nie, M., Chen, P., Zhao, Y.-n., Shen, J., Wang, C., Xiong, Y., Yin, K., and Sun, L. Wearable multimode sensor with a seamless integrated structure for recognition of different joint motion states with the assistance of a deep learning algorithm. *Microsystems & nanoengineering*, 8 (1):1–14, 2022.

Wu, N., Jastrzebski, S., Cho, K., and Geras, K. J. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, 2022.

Xiao, Y., Codevilla, F., Gurram, A., Urfalioglu, O., and López, A. M. Multimodal end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

Xie, Z., Wang, S. I., Li, J., Lévy, D., Nie, A., Jurafsky, D., and Ng, A. Y. Data noising as smoothing in neural network language models. In *International Conference on Learning Representations*, 2017.

Yadav, A. and Vishwakarma, D. K. A deep multi-level attentive network for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(1), 2023.

Zhang, L., Zhu, X., Chen, X., Yang, X., Lei, Z., and Liu, Z. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

# Appendix

## A. Proofs

### A.1. Proof of Theorem 1

*Proof.* Let $(x, y) \sim \mathcal{D}$ denotes the multimodal sample, then we have

$$\ell(f(x), y) = \ell(\sum_{m=1}^{M} w^m f^m(x^{(m)}), y). \tag{19}$$

Noted that $\ell$ is a convex logistic loss function, which indicates that

$$\ell(f(x), y) = \ell(\sum_{m=1}^{M} w^m f^m(x^{(m)}), y) \leq \sum_{m=1}^{M} w^m \ell(f^m(x^{(m)}), y). \tag{20}$$

Then we take the expectation on both sides of the above equation

$$\mathbb{E}_{(x,y)\sim\mathcal{D}} \ell(f(x), y) \leq \mathbb{E}_{(x,y)\sim\mathcal{D}} \sum_{m=1}^{M} w^m \ell(f^m(x^{(m)}), y), \tag{21}$$

since expectation is a linear operator and the expected value of the product is equal to the product of the expected values plus the covariance, we can further decompose the right-hand side of the equation into

$$\mathbb{E}_{(x,y)\sim\mathcal{D}} \ell(f, y) \leq \sum_{m=1}^{M} \mathbb{E}_{(x,y)\sim\mathcal{D}}[w^m \ell(f^m, y)] \tag{22}$$

$$= \sum_{m=1}^{M} \mathbb{E}_{(x,y)\sim\mathcal{D}}(w^m) \mathbb{E}_{(x,y)\sim\mathcal{D}}(\ell(f^m, y)) + Cov(w^m, \ell(f^m, y)) \tag{23}$$

Next, we recap the Rademacher complexity measure for model complexity. We use complexity-based learning theory (Bartlett & Mendelson, 2002) (Theorem 8) to quantify the generalization error of unimodal models.

Let $D_{\text{train}} = \{x_i, y_i\}_{i=1}^{N}$ be the training dataset of $N$ samples, $\hat{E}(f^m)$ is the unimodal empirical error of $f^m$. Then for any hypothesis $f^m$ in $\mathcal{H}$ (i.e., $\mathcal{H} : \mathcal{X} \to \{-1, 1\}$, $f \in \mathcal{H}$) and $1 > \delta > 0$, with probability at least $1 - \delta$, we have

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}(f^m) \leq \hat{E}(f^m) + \mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{ln(1/\delta)}{2N}},$$

where $\mathfrak{R}_m(f^m)$ is the Rademacher complexities.

Finally, it holds with probability at least $1 - \delta$ that

$$\text{GError}(f) \leq \sum_{m=1}^{M} \mathbb{E}(w^m) \hat{E}(f^m) + \mathbb{E}(w^m) \mathfrak{R}_m(\mathcal{H}) + Cov(w^m, \ell(f^m, y)) + M\sqrt{\frac{ln(1/\delta)}{2N}}. \tag{24}$$

□

### A.2. Proof of Theorem 2

*Proof.* Let $\mathcal{O}(\text{GError}(f_{\text{dynamic}}))$, $\mathcal{O}(\text{GError}(f_{\text{static}}))$ be the upper bound of generalization error of multimodal classifier using dynamic and static fusion strategy respectively, $\hat{E}(f^m)$ is the unimodal empirical errors of $f^m$ on $D_{\text{train}}$ defined in Theorem. 1. Theoretically, optimizing over the same function class results in the same empirical risk. Therefore

$$\hat{E}(f^m_{\text{static}}) = \hat{E}(f^m_{\text{dynamic}}). \tag{25}$$

Additionally, the intrinsic complexity of unimodal classifier $\mathfrak{R}_m(f^m)$ is also invariant

$$\mathfrak{R}_m(f^m_{\text{static}}) = \mathfrak{R}_m(f^m_{\text{dynamic}}). \tag{26}$$

Thus in this special case, it holds that

$$\sum_{m=1}^{M} \mathbb{E}(w^m_{\text{dynamic}})\hat{E}(f^m) \leq \sum_{m=1}^{M} w^m_{\text{static}}\hat{E}(f^m), \tag{27}$$

and

$$\sum_{m=1}^{M} \mathbb{E}(w^m_{\text{dynamic}})\mathfrak{R}_m(f^m) \leq \sum_{m=1}^{M} w^m_{\text{static}}\mathfrak{R}_m(f^m), \tag{28}$$

if $\mathbb{E}(w^m_{\text{dynamic}}) = w^m_{\text{static}}$.

Since the covariance and correlation coefficient have the same sign, when $r(w^m, l^m) \leq 0$, the covariance $Cov(w^m, l^m)$ is also less than or equal to 0. Therefore, it holds that

$$\mathcal{O}(\text{GError}(f_{\text{dynamic}})) \leq \mathcal{O}(\text{GError}(f_{\text{static}})) \tag{29}$$

with probability at least $1 - \delta$, if we have

$$\mathbb{E}(w^m_{\text{dynamic}}) = w^m_{\text{static}} \tag{30}$$

and

$$r(w^m_{\text{dynamic}}, \ell(f^m)) \leq 0 \tag{31}$$

for all input modality $m$. □

## B. Experimental details

### B.1. Datasets details

∘ **Senses recognition.** For NYUD-V2, following the standard split, we reorganize the 27 categories into 10 categories with 9 usual senes and one "others" category. For SUN RGB-D, following the previous work (Han et al., 2021), we use the 19 major scene categories of SUN RGB-D, each of which contains at least 80 images.
∘ **Image-text classification.** For FOOD-101, following the previous work (Kiela et al., 2019), there are 60101 image-text pairs in the training set, 5000 image-text pairs in the validation set, and 21695 image-text pairs in the test set. For MVSA, we conduct the division strategy presents in (Kiela et al., 2019). There are 1555 image-text pairs in the training set. The validation set contains 518 image-text pairs, and the test set contains 519 image-text pairs.

### B.2. Implementation details

**Senses recognition.** For senses recognition task, we compare the proposed method with diverse multimodal fusion methods, including late fusion, align-based fusion, concatenated-based fusion and recent SOTA MMTM (attention-based fusion). Regarding late fusion, align-based fusion, concatenated-based fusion, we adopt the architecture of ResNet (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) as the backbone network for each modality. ∘ **Concatenate-based fusion** For concatenate-based fusion, we concatenate the representations extracted from different modalities by ResNet. Then a fully connection layer is deployed to map the multimodal representation to the target space. The dimensions of unimodal representation and common representation are 128 and 256 respectively. ∘ **Align-based fusion** The alignment fusion method is a re-implementation of (Wang et al., 2016). We deploy cosine distance to measure the similarity of representations. ∘ **MMTM** We follow the authors' implementation, where the squeeze ratio is set to 4. For all compared methods, we use Adam optimizer to train all above models with $\mathcal{L}_2$ regularization and dropout (Srivastava et al., 2014). The learning rate is 1e-4 and the dropout rate is 0.1.

**Image-text classification.** For image-text classification, we compare the proposed method with diverse multimodal fusion methods, including late fusion, concatenated-bow fusion, concatenated-bert fusion and MMTM. For late fusion and concatenated-bert fusion, we adopt the architecutre of ResNet (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) as

the backbone network for image modality and pre-trained Bert(Devlin et al., 2018) for text modality. For concatenated-Bow fusion, we use the Bow (Pennington et al., 2014) to replace BERT for text modality. For the Bert models, we use BertAdam and regular Adam for the other models. The learning rate is 1e-4 with a warmup rate of 0.1. We adopt the early stop strategy based on validation accuracy.

For all above experiments, we conduct sampling during the whole training phase ($\mathrm{T_s} = 1$). The hyperparameter $\lambda$ is set to $0.1$. Temperature parameters $\{\mathcal{T}^m\}_{m=1}^{M}$ are set to $1$.

## C. Additional results

### C.1. Full results with standard deviation

In this section, we present the full results with standard deviation in Tab. 5, and Tab. 4.

### C.2. Different type of noise

We provide more results with different type of noise (i.e., salt-pepper noise with varying noise rate $\epsilon$) in Tab. 6. The results validate that the proposed method can improve the performance of multimodal fusion methods under different type of noise.

*Table 4.* Full ablation study on NYU Depth V2.

| UAW | $\mathcal{L}_{\text{reg}}$ | $\epsilon = \mathbf{0.0}$ | $\epsilon = \mathbf{5.0}$ | $\epsilon = \mathbf{10.0}$ | $\epsilon = \mathbf{20.0}$ |
|---|---|---|---|---|---|
| ✗ | ✗ | $69.14 \pm 0.69$ | $68.35 \pm 0.82$ | $59.62 \pm 1.17$ | $53.98 \pm 1.08$ |
| ✗ | ✓ | $69.68 \pm 0.39$ | $67.74 \pm 0.40$ | $61.35 \pm 0.34$ | $58.26 \pm 0.1.13$ |
| ✓ | ✗ | $70.06 \pm 0.1.03$ | $\mathbf{69.11 \pm 0.82}$ | $61.59 \pm 0.0.72$ | $57.49 \pm 1.41$ |
| ✓ | ✓ | $\mathbf{70.09 \pm 0.38}$ | $68.81 \pm 0.62$ | $\mathbf{61.62 \pm 0.31}$ | $\mathbf{58.87 \pm 0.40}$ |

*Table 5.* Full comparison results when 50% of the modalities are corrupted with Gaussian noise.

| Dataset | Dynamic | Method | $\epsilon = \mathbf{0.0}$ | $\epsilon = \mathbf{5.0}$ | $\epsilon = \mathbf{10.0}$ |
|---|---|---|---|---|---|
| NYU Depth V2 | ✗ | RGB | $62.65 \pm 1.22$ | $50.95 \pm 3.38$ | $44.13 \pm 3.80$ |
| | ✗ | Depth | $63.30 \pm 0.48$ | $53.12 \pm 1.52$ | $45.46 \pm 2.07$ |
| | ✗ | Late fusion | $69.14 \pm 0.67$ | $59.63 \pm 2.44$ | $51.99 \pm 3.11$ |
| | ✗ | Concat | $70.31 \pm 0.80$ | $59.97 \pm 2.14$ | $53.20 \pm 3.50$ |
| | ✗ | Align | $70.31 \pm 1.28$ | $59.47 \pm 1.84$ | $51.74 \pm 3.41$ |
| | ✓ | MMTM | $71.04 \pm 0.41$ | $60.37 \pm 2.61$ | $52.28 \pm 3.77$ |
| | ✓ | TMC | $\mathbf{71.06 \pm 0.76}$ | $61.04 \pm 1.66$ | $53.36 \pm 2.76$ |
| | ✓ | Ours | $70.09 \pm 0.97$ | $\mathbf{61.62 \pm 1.84}$ | $\mathbf{55.60 \pm 2.09}$ |
| SUN RGB-D | ✗ | RGB | $52.99 \pm 0.88$ | $37.81 \pm 1.14$ | $33.07 \pm 1.81$ |
| | ✗ | Depth | $56.78 \pm 0.19$ | $48.40 \pm 1.11$ | $42.94 \pm 1.63$ |
| | ✗ | Late fusion | $62.00 \pm 0.15$ | $52.52 \pm 0.67$ | $47.48 \pm 1.40$ |
| | ✗ | Concat | $\mathbf{62.48 \pm 0.50}$ | $53.30 \pm 0.39$ | $48.01 \pm 0.96$ |
| | ✗ | Align | $61.12 \pm 0.61$ | $50.05 \pm 1.59$ | $44.19 \pm 2.18$ |
| | ✓ | MMTM | $61.72 \pm 0.67$ | $51.86 \pm 1.14$ | $46.03 \pm 1.47$ |
| | ✓ | TMC | $60.68 \pm 0.24$ | $51.24 \pm 0.96$ | $45.66 \pm 2.06$ |
| | ✓ | Ours | $62.09 \pm 0.56$ | $\mathbf{53.40 \pm 0.89}$ | $\mathbf{48.58 \pm 0.82}$ |
| UMPC FOOD101 | ✗ | Bow | $82.50 \pm 0.18$ | $61.68 \pm 0.71$ | $41.95 \pm 0.54$ |
| | ✗ | Img | $64.62 \pm 0.40$ | $34.72 \pm 0.53$ | $33.03 \pm 0.37$ |
| | ✗ | Bert | $86.46 \pm 0.05$ | $67.38 \pm 0.19$ | $43.88 \pm 0.32$ |
| | ✗ | Late fusion | $90.69 \pm 0.12$ | $68.49 \pm 3.37$ | $57.99 \pm 1.59$ |
| | ✗ | Concatbow | $70.77 \pm 0.09$ | $38.28 \pm 0.26$ | $35.68 \pm 0.69$ |
| | ✗ | Concatbert | $88.20 \pm 0.34$ | $61.10 \pm 2.02$ | $49.86 \pm 2.05$ |
| | ✓ | MMBT | $91.52 \pm 0.10$ | $72.32 \pm 0.34$ | $56.75 \pm 0.33$ |
| | ✓ | TMC | $89.86 \pm 0.07$ | $73.93 \pm 0.34$ | $61.37 \pm 0.21$ |
| | ✓ | Ours | $\mathbf{92.92 \pm 0.11}$ | $\mathbf{76.03 \pm 0.70}$ | $\mathbf{62.21 \pm 0.25}$ |
| MVSA | ✗ | Bow | $48.79 \pm 7.05$ | $42.20 \pm 6.40$ | $41.57 \pm 6.28$ |
| | ✗ | Img | $64.12 \pm 1.23$ | $49.36 \pm 2.02$ | $45.00 \pm 2.63$ |
| | ✗ | Bert | $75.61 \pm 0.53$ | $69.50 \pm 1.50$ | $47.41 \pm 0.79$ |
| | ✗ | Late fusion | $76.88 \pm 1.30$ | $63.46 \pm 3.46$ | $55.16 \pm 3.60$ |
| | ✗ | ConcatBow | $64.08 \pm 1.54$ | $49.95 \pm 2.29$ | $45.39 \pm 3.03$ |
| | ✗ | ConcatBert | $65.59 \pm 1.33$ | $50.70 \pm 2.65$ | $46.12 \pm 2.44$ |
| | ✓ | MMBT | $78.50 \pm 0.40$ | $71.99 \pm 1.04$ | $55.34 \pm 2.84$ |
| | ✓ | TMC | $74.87 \pm 2.24$ | $66.72 \pm 4.55$ | $60.35 \pm 2.79$ |
| | ✓ | Ours | $\mathbf{78.07 \pm 1.10}$ | $\mathbf{73.85 \pm 1.42}$ | $\mathbf{61.28 \pm 2.12}$ |

*Table 6.* Full comparison results when 50% of the modalities are corrupted with Salt-pepper noise.

| Dataset | Dynamic | Method | $\epsilon = \mathbf{0.0}$ | $\epsilon = \mathbf{5.0}$ | $\epsilon = \mathbf{10.0}$ |
|---|---|---|---|---|---|
| NYU Depth V2 | ✗ | RGB | $62.61 \pm 1.21$ | $49.14 \pm 1.40$ | $34.76 \pm 1.59$ |
| | ✗ | Depth | $63.32 \pm 0.50$ | $50.99 \pm 1.41$ | $38.56 \pm 2.16$ |
| | ✗ | Late fusion | $69.16 \pm 0.68$ | $56.27 \pm 2.40$ | $41.22 \pm 2.78$ |
| | ✗ | Concat | $70.44 \pm 0.89$ | $57.98 \pm 2.08$ | $44.51 \pm 2.91$ |
| | ✗ | Align | $70.31 \pm 1.28$ | $57.54 \pm 2.50$ | $43.01 \pm 2.66$ |
| | ✓ | MMTM | $\mathbf{71.04 \pm 0.41}$ | $\mathbf{59.45 \pm 1.38}$ | $44.59 \pm 2.49$ |
| | ✓ | TMC | $71.01 \pm 0.75$ | $59.34 \pm 1.03$ | $44.65 \pm 2.30$ |
| | ✓ | Ours | $70.06 \pm 0.81$ | $58.50 \pm 2.05$ | $\mathbf{45.69 \pm 2.79}$ |
| SUN RGB-D | ✗ | RGB | $52.63 \pm 0.89$ | $40.42 \pm 0.99$ | $28.15 \pm 1.00$ |
| | ✗ | Depth | $56.81 \pm 0.57$ | $46.36 \pm 0.82$ | $35.66 \pm 1.44$ |
| | ✗ | Late fusion | $61.79 \pm 0.57$ | $51.54 \pm 2.12$ | $39.35 \pm 2.89$ |
| | ✗ | Concat | $\mathbf{62.06 \pm 0.53}$ | $51.09 \pm 1.91$ | $38.61 \pm 3.07$ |
| | ✗ | Align | $61.02 \pm 0.54$ | $50.45 \pm 0.82$ | $38.70 \pm 1.46$ |
| | ✓ | MMTM | $61.80 \pm 0.40$ | $51.09 \pm 0.77$ | $38.38 \pm 1.56$ |
| | ✓ | TMC | $61.02 \pm 0.39$ | $50.88 \pm 1.28$ | $39.61 \pm 2.30$ |
| | ✓ | Ours | $61.89 \pm 0.49$ | $\mathbf{52.49 \pm 1.81}$ | $\mathbf{40.53 \pm 2.79}$ |
| UMPC FOOD101 | ✗ | Bow | $82.43 \pm 0.18$ | $60.83 \pm 0.54$ | $41.56 \pm 0.33$ |
| | ✗ | Img | $64.53 \pm 0.47$ | $50.75 \pm 0.44$ | $36.83 \pm 0.92$ |
| | ✗ | Bert | $86.44 \pm 0.02$ | $67.41 \pm 0.20$ | $43.89 \pm 0.33$ |
| | ✗ | Late fusion | $90.66 \pm 0.16$ | $77.99 \pm 0.54$ | $58.75 \pm 0.99$ |
| | ✗ | Concatbow | $70.68 \pm 0.12$ | $55.01 \pm 0.33$ | $38.81 \pm 0.62$ |
| | ✗ | Concatbert | $88.22 \pm 0.36$ | $72.49 \pm 0.75$ | $52.10 \pm 0.97$ |
| | ✓ | MMBT | $91.51 \pm 0.10$ | $76.27 \pm 0.22$ | $54.98 \pm 0.55$ |
| | ✓ | TMC | $89.86 \pm 0.07$ | $77.86 \pm 0.41$ | $60.22 \pm 0.43$ |
| | ✓ | Ours | $\mathbf{92.90 \pm 0.13}$ | $\mathbf{80.87 \pm 0.40}$ | $\mathbf{61.60 \pm 0.20}$ |
| MVSA | ✗ | Bow | $48.82 \pm 7.08$ | $42.23 \pm 6.43$ | $41.60 \pm 6.31$ |
| | ✗ | Img | $64.12 \pm 1.23$ | $56.72 \pm 1.92$ | $50.71 \pm 3.20$ |
| | ✗ | Bert | $75.61 \pm 0.53$ | $69.50 \pm 1.50$ | $47.41 \pm 0.79$ |
| | ✗ | Late fusion | $76.88 \pm 1.30$ | $67.88 \pm 1.87$ | $55.43 \pm 1.94$ |
| | ✗ | ConcatBow | $64.08 \pm 1.54$ | $56.66 \pm 1.73$ | $49.35 \pm 2.44$ |
| | ✗ | ConcatBert | $65.59 \pm 1.33$ | $58.69 \pm 2.25$ | $51.16 \pm 2.99$ |
| | ✓ | MMBT | $\mathbf{78.50 \pm 0.40}$ | $\mathbf{74.07 \pm 1.12}$ | $51.26 \pm 5.65$ |
| | ✓ | TMC | $74.87 \pm 2.24$ | $68.02 \pm 3.07$ | $56.62 \pm 3.67$ |
| | ✓ | Ours | $78.07 \pm 1.10$ | $73.90 \pm 1.89$ | $\mathbf{60.41 \pm 2.63}$ |