Embracing Unimodal Aleatoric Uncertainty for Robust Multimodal Fusion

Zixian Gao[1]*, Xun Jiang[1]*, Xing Xu[1,2]†, Fumin Shen[1], Yujie Li[3], Heng Tao Shen[1,2]
[1]Center for Future Media & School of Computer Science and Engineering,
University of Electronic Science and Technology of China, China
[2]College of Electronic and Information Engineering, Tongji University, China
[3]Kyushu Institute of Technology, Japan

Abstract

*As a fundamental problem in multimodal learning, multimodal fusion aims to compensate for the inherent limitations of a single modality. One challenge of multimodal fusion is that the unimodal data in their unique embedding space mostly contains potential noise, which leads to corrupted cross-modal interactions. However, in this paper, we show that the potential noise in unimodal data could be well quantified and further employed to enhance more stable unimodal embeddings via contrastive learning. Specifically, we propose a novel generic and robust multimodal fusion strategy, termed **E**mbracing **A**leatoric **U**ncertainty (**EAU**), which is simple and can be applied to kinds of modalities. It consists of two key steps: (1) the Stable Unimodal Feature Augmentation (SUFA) that learns a stable unimodal representation by incorporating the aleatoric uncertainty into self-supervised contrastive learning. (2) Robust Multimodal Feature Integration (RMFI) leveraging an information-theoretic strategy to learn a robust compact joint representation. We evaluate our proposed EAU method on five multimodal datasets, where the video, RGB image, text, audio, and depth image are involved. Extensive experiments demonstrate the EAU method is more noise-resistant than existing multimodal fusion strategies and establishes new state-of-the-art on several benchmarks.*

1. Introduction

By exploring the complementary information from different modalities, multimodal learning has achieved impressive success in kinds of artificial intelligence applications, such as multimodal image classification in Internet applications [1–5] and sentiment analysis in intelligent robotics [6–9]. It has been widely proved that fusing information from different modalities appropriately is helpful to gain better performance. Dedicated to this, multimodal fusion has become an emerging challenge in multimodal learning, which integrates different modalities into a unified representation with powerful neural networks.

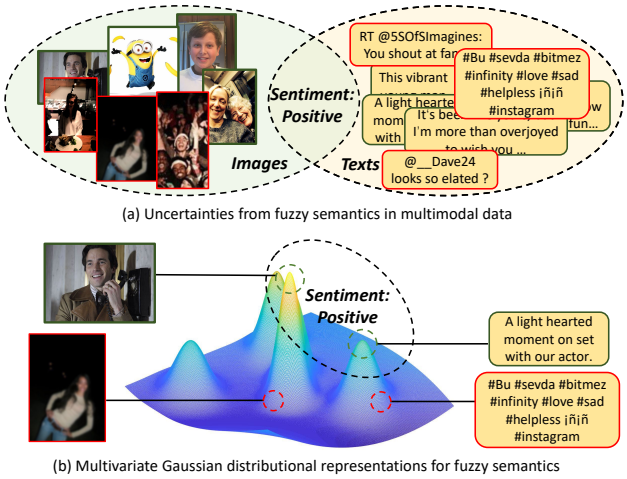

Figure 1. An illustration of the aleatoric uncertainty in multimodal datasets and the distributional representations: (a) As the semantics are ambiguous, the multimodal data are prone to introducing noisy data. (b) We adopt a multivariate Gaussian distribution to represent the fuzzy semantics in a noisy latent space.

Nevertheless, with the studies developing, researchers have recognized that the multimodal data in different modal forms may be unreliable due to the unique noise in their own modal space. Recently, several works [5, 10–13] show that previous multimodal fusion methods have overlooked the unreliable quality of multimodal data. Specifically, the widely adopted fusion strategies may fail on noisy multimodal data because the cross-modal interactions may obtain limited influence by the uncertainty in data. Here we illustrate typical noise in multimodal data in Fig. 1. Due to the semantics of "Positive" being fuzzy and the labels being judged by humans subjectively, the given image-text pairs are all recognized as positive in sentiment analysis, even though there is obvious noise in both image and text

*These authors contributed equally to this work.
†Corresponding author.

modalities. In general, we can say the introduced noise comes from the aleatoric uncertainty [14], which undermines the effectiveness of exploiting multimodal data and demonstrates multimodal learning is not a free lunch.

Hence, with the emerging challenge of uncertainty in multimodal data, we raise the first fundamental question: *Can we quantify the uncertainty in multimodal data?* Inspired by probability Distributional Representation [15–18], we naturally adopt the Gaussian distributions to work out this problem. We assume that each instance can be represented as a multivariate normal distribution, where the variance can be regarded as the intrinsic aleatoric uncertainty. With the quantified uncertainty, we are allowed to take a closer look at the multimodal fusion and raise the second question: *Is it appropriate to completely drop the intrinsic uncertainty?* From the examples in Fig. 1(a), we can observe that even if image-text pairs reveal similar semantics, the aleatoric uncertainty is still unavoidable because of the domain shift, extra descriptions, or image quality, *etc*. To this end, as is illustrated in Fig. 1(b), we argue that a multivariate normal distribution that considers aleatoric uncertainty can be regarded as a fuzzy representation for semantics, where the semantic-related data are within similar distributions, even if they are in different modalities. Motivated by the two assumptions, we develop a novel multimodal fusion strategy in this paper, *i.e.*, Embracing Aleatoric Uncertainty (EAU).

Specifically, our EAU method consists of the following two processes: (1) The Stable Unimodal Feature Augmentation (SUFA) quantifies the intrinsic aleatoric uncertainty in each modality by representing samples as multivariate normal distributions. Then, we further conduct self-supervised contrastive learning with these distributional representations to learn stable unimodal embeddings with better semantical consistency. (2) Robust Multimodal Feature Integration (RMFI) that dynamically fuses the stable unimodal embeddings into a joint representation. Particularly, considering that SUFA focuses on semantic consistency only and ignores the problem of information redundancy, we employ an information-theoretic strategy, *i.e.*, Variational Information Bottleneck [19, 20], to learn a compact joint representation with less redundancy. We evaluate our method on five multimodal benchmark datasets and demonstrate it outperforms existing state-of-the-art methods on both Multimodal Sentiment Analysis and Multimodal Image Classification tasks. Moreover, our proposed EAU method also reveals better robustness on noisy datasets compared with the counterpart methods [2, 3, 5].

Overall, our contributions in this paper can be summarized as three-fold:

- We propose a novel multimodal fusion method termed Embracing Aleatoric Uncertainty (EAU), which quantifies the intrinsic aleatoric uncertainty and leverages it to learn stable and robust joint representations.
- We devise the Stable Unimodal Feature Augmentation (SUFA) module, which quantifies aleatoric uncertainty and learns stable unimodal representations with better semantical consistency.
- We design the Robust Multimodal Feature Integration (RMFI) module to learn compact joint representations with less redundancy, which further improves the robustness of multimodal fusion in our method.

2. Related Work

Multimodal Fusion. Multimodal fusion, which aims at learning stronger representations from different modalities, has become an essential part of a spectrum of computer vision research, such as vision-audio learning [21, 22], vision-language learning [23, 24], image retrieval [25–27], and video understanding [28–30]. Generally, multimodal fusion can be categorized into three types: early fusion, intermediate fusion, and late fusion. The previous multimodal fusion methods can be categorized into early fusion and late fusion in terms of their feature-level or decision-level fusion operations. In the last decades, massive works with deep learning suggest that intermediate fusion [13, 29, 31], which learns unified embeddings for multimodal data in the hidden layers of neural networks, could benefit representation learning. Although these multimodal fusion methods achieve remarkable improvements in kinds of multimodal learning tasks, most of them ignore the uncertainty in multimodal learning. Recently, several works [5, 10–13] empirically or theoretically demonstrated that conventional multimodal fusion methods had limited performance and robustness on noisy or corrupted multimodal data. Inspired by these pioneering works, we devised a novel intermediate multimodal fusion method, which could quantify the uncertainty and learn more robust joint representations.

Uncertainty in Deep Learning. In general, the uncertainties in deep learning can be categorized into epistemic uncertainty and aleatoric uncertainty [14]. The former aims at capturing the noise of the parameters in deep neural networks, while the latter measures the noise inherent in given training data. To improve the robustness and generalization of open-world scenarios, many researchers incorporate the uncertainty estimation into the deep learning models for computer vision tasks, such as face recognition [15, 16], semantic segmentation [32], and action localization [33, 34]. There is also a list of works that incorporate uncertainty estimation into multimodal learning tasks [5, 17, 22, 35]. However, most of them only quantified uncertainties to learn probability distribution representations that are used for cross-modal interactions, but ignore the value of intrinsic uncertainties and hardly take a further analysis on the robustness of multimodal fusion. Following this observation,

we develop this work to learn more robust compact joint representations for multimodal learning, which embraces the uncertainties via self-supervised contrastive learning.

Information Bottleneck Theory. The Information Bottleneck Theory [19] was initially formulated in the context of signal processing, which is proposed to discover a more concise representation of a signal while retaining its utmost informative content. Alemi *et al*. [36] proposed the Variational Information Bottleneck (VIB) to bridge the gap between Information Bottleneck Theory and deep learning, which approximates the information bottleneck constraints and enables it in deep learning. Based on the VIB, massive researchers introduce the Information Bottleneck Theory in deep learning models to tackle kinds of computer vision tasks, such as object detection [37–39] or image classification [37,38]. Furthermore, as the VIB aims at learning compact minimal representations, it also attracts wide attention from the multimodal learning field [28, 31]. For example, Mai *et al*. [28] devised a multimodal information bottleneck to learn the minimal sufficient unimodal and multimodal representations. Inspired by their successes, we also devise an multimodal integration strategy based on VIB to reduce the redundancy in joint representations, which further improves the robustness and effectiveness.

3. Proposed Method

3.1. Preliminaries

Uncertainty Estimation in Deep Learning Models. In terms of [14], the uncertainties in deep learning models can be categorized into aleatoric uncertainty and epistemic uncertainty. The former refers to the uncertainty inherent in the observations and can not be explained away with more data, while the latter only accounts for uncertainty in the model and can be explained away given enough data. Typically, given a deep learning model $f_\theta(\cdot)$, it conducts the mapping $\mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ represent input data and observed label space respectively. The aleatoric uncertainty in $\mathcal{X}$ will lead to a corrupted predicted results in $\mathcal{Y}$. To estimate the uncertainties in deep learning models, a widely adopted strategy is sampling N weights $\theta \sim p(\theta|x, y), x \in \mathcal{X}, y \in \mathcal{Y}$ via Dropout operations for different predictions for mean μ and variance σ^2, and adopting a Gaussian posterior $p(y|x) \sim \mathcal{N}(\mu_*(x), \sigma_*^2(x))$. Here we can get the predictive variance and extract the aleatoric uncertainty which represents the ambiguity in x:

$$\begin{aligned} \mu_*^2(x) &= \frac{1}{N}\sum\nolimits_i \mu_i^2(x), \\ \sigma_*^2(x) &= \frac{1}{N}\sum\nolimits_i \sigma_i^2(x) + \frac{1}{N}\sum\nolimits_i \mu_i^2(x) - \mu_*^2(x) \\ &= \mathbb{E}_i\left[\sigma_i^2(x)\right] + \mathrm{Var}_i\left[\mu_i(x)\right], \end{aligned} \tag{1}$$

where $\mathbb{E}_i\left[\sigma_i^2(x)\right]$ represents the aleatoric uncertainty, while $\mathrm{Var}_i\left[\mu_i(x)\right]$ is the epistemic uncertainty, which attributes to the model $f_\theta(\cdot)$ rather than input data $\mathcal{X}$. In our work, we pay more attention to the aleatoric uncertainty since our goal is to quantify the intrinsic noise in multimodal data and leverage it for learning better joint representations.

Variational Information Bottleneck. The Variational Information Bottleneck (VIB) [20] is an information-theoretic strategy widely adopted in deep learning models, which aims to maintain maximal discriminative feature representations with minimal redundancy. Specifically, given the input variable $\mathbf{x}$ with noise or redundant information and the target variable $\mathbf{y}$, the VIB is to learn the compressed latent variable $\mathbf{z}$, and $\mathbf{z}$ is maximally discriminative about the target variable $\mathbf{y}$. Moreover, since the input $\mathbf{x}$ is noisy and redundant, the VIB also requires $\mathbf{z}$ to be minimally discriminative about the original variable $\mathbf{x}$. In our work, we leverage the VIB to learn compact joint representations, which overcome the redundancy brought by highly aligned distributional multimodal representations to improve the robustness of multimodal fusion.

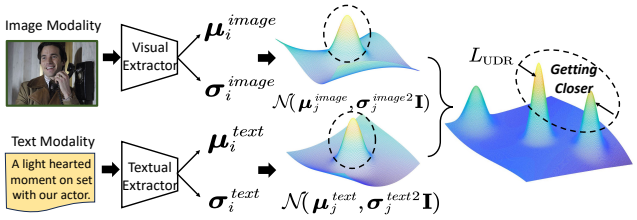

Figure 2. An illustration of the Unimodal Distributional Representation process. For clarity, we visualize the multivariate Gaussian distributions with two variables only.

3.2. Stable Unimodal Feature Augmentation

As the data in different modalities contain unique noise in their own modal space, we first propose the Stable Unimodal Feature Augmentation (SUFA) module to quantify their intrinsic aleatoric uncertainty. Following the argument that the uncertainty is attributed to the fuzzy representation of semantics, we further leverage the aleatoric uncertainty via self-supervised contrastive learning.

Unimodal Distributional Representation. Given the multimodal samples $\mathbf{x}_i^m, m \in \mathcal{M}$ where $\mathcal{M}$ is the modality set including images, audio, text, *etc*., we learn the distributional representations to quantify the aleatoric uncertainty in each modality. According to the deviation listed in Eq. 1, we can observe that the aleatoric uncertainty can be predicted with deep learning models $f_\theta(\cdot)$ as the variance $\boldsymbol{\sigma}_i^{m2}$ directly if the epistemic uncertainty is not considered. To this end, we first employ the corresponding feature extractor to learn the preliminary embeddings for each modality, then deploy two additional fully connected layers to learn a mean vector $\boldsymbol{\mu}_i^{m2} \in \mathbb{R}^d$ and a variance vector $\boldsymbol{\sigma}_i^{m2} \in \mathbb{R}^d$.

Furtherly, we define the representation $\mathbf{z}_i^m$ in latent space of each sample $\mathbf{x}_i^m$ as a multivariate Gaussian distribution with d variable, which can be represented as:

$$\begin{aligned} p(\mathbf{z}_i^m|\mathbf{x}_i^m) &\sim \mathcal{N}(\boldsymbol{\mu}_i^m, \boldsymbol{\sigma}_i^{m2}\mathbf{I}), \\ \boldsymbol{\mu}_i^m = f_{\theta_1^m}(\mathbf{x}_i^m), &\ \boldsymbol{\sigma}_i^m = f_{\theta_2^m}(\mathbf{x}_i^m), \end{aligned} \tag{2}$$

where $f_{\theta_1^m}(\cdot)$ and $f_{\theta_2^m}(\cdot)$ represent the two different fully connected layers for mean and variance respectively. $\mathbf{I}$ is the identity matrix. To maintain the consistency of semantics in different modalities, we further align the distributional representations of a multimodal sample with the Kullback-Leibler divergence:

$$\begin{aligned} L_{\text{UDR}} = &\sum_{m_1 \neq m_2}^{m_1, m_2 \in \mathcal{M}} KL\left(p(\mathbf{z}_i^{m_1}|\mathbf{x}_i^{m_1}) \| p(\mathbf{z}_i^{m_2}|\mathbf{x}_i^{m_2})\right) \\ &+ \sum_{m}^{m \in \mathcal{M}} KL\left(p(\mathbf{z}_i^m|\mathbf{x}_i^m) \| \mathcal{N}(0, \mathbf{I})\right). \end{aligned} \tag{3}$$

We illustrate the Unimodal Distributional Representation process in Fig. 2 with bi-modal input for clarity. It can be observed the L_{UDR} will push two multimodal distributional representations with similar semantics closer. In this way, the representation of each multimodal sample is not limited within a deterministic point embedding, but a consistent fuzzy representation on several multivariate Gaussian distributions. Particularly, the variance $\boldsymbol{\sigma}_i^m$ reveals the aleatoric uncertainty in m modality and the mean $\boldsymbol{\mu}_i^m$ is the corresponding stable representation.

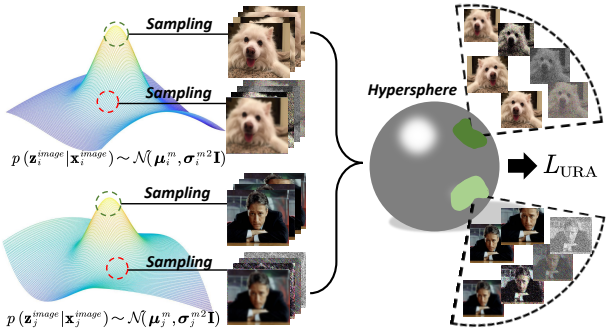

Figure 3. Illustration of the Uncertainty-based Representation Augmentation process. Here we show the distributional representations of image modality only for clarity.

Uncertainty-based Representation Augmentation. With the quantified aleatoric uncertainty, we consider the second question we raised in Sec.1: *should we drop the aleatoric uncertainty in multimodal data?* Intuitively, the aleatoric uncertainty in multimodal data is unavoidable since the natural ambiguity of semantics. However, it also results in the diversity of unimodal data in different modalities. To this end, we leverage the aleatoric uncertainty to generate the unseen samples such that learned unimodal representations are insensitive to diverse unimodal input with similar semantics. Take the image modality as an example, we illustrate the Uncertainty-based Representation Augmentation process in Fig. 3. Specifically, given a unimodal distributional representation $p(\mathbf{z}_i^m|\mathbf{x}_i^m) \sim \mathcal{N}(\boldsymbol{\mu}_i^m, \boldsymbol{\sigma}_i^{m2}\mathbf{I})$, we first random sample an anchor point $\tilde{\mathbf{z}}_i^m$ and an augmented point $\mathbf{z}_i^m$ from the multivariate Gaussian distribution as matching pairs. Moreover, we randomly sample a set of negative points from other distributional representations and devise a self-supervised contrastive learning mechanism as follows:

$$L_{\text{URA}} = -\log \frac{e^{\text{sim}(\tilde{\mathbf{z}}_i^m, \mathbf{z}_i^m)/\tau}}{\sum_{j \neq i}\left(e^{\text{sim}(\tilde{\mathbf{z}}_i^m, \mathbf{z}_i^m)/\tau} + e^{\text{sim}(\tilde{\mathbf{z}}_i^m, \mathbf{z}_j^m)/\tau}\right)}, \tag{4}$$

where τ is a temperature factor and sim is cosine similarity calculation. Here we adopt the re-parameterization trick [40] to conduct the sampling operations, which can be formulated as follows:

$$\mathbf{z}_i^m = \boldsymbol{\mu}_i^m + \epsilon \boldsymbol{\sigma}_i^m, \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{5}$$

3.3. Robust Multimodal Feature Integration

With the SUFA, we obtain stable unimodal representations with consistent semantics. However, we only consider the consistency in each modality but ignore the redundancy caused by duplicated representations. To this end, we proposed the Robust Multimodal Feature Integration (RMFI) module in this section. The overview of the RMFI module is illustrated in Fig. 4.

Dynamic Multimodal Integration. Inspired by dynamic multimodal fusion [5], we assume that different modalities have unequal contributions to the joint representations in terms of the observed label space. To this end, we first apply a Dynamic Multimodal Integration strategy based on an attention mechanism. Specifically, given the stable unimodal representations $\boldsymbol{\mu}_i^m$, we calculate the attentive weights in the distributional representations across modalities according to their quantified uncertainties $\boldsymbol{\sigma}_i^m$, and apply them for multimodal integration:

$$\hat{\mathbf{x}}_i = \sum_{m \in \mathcal{M}} \alpha_i^m \boldsymbol{\mu}_i^m, \quad \alpha_i^m = \frac{e^{\frac{1}{\overline{\boldsymbol{\sigma}}_i^m}}}{\sum_{m \in \mathcal{M}} e^{\frac{1}{\overline{\boldsymbol{\sigma}}_i^m}}}, \tag{6}$$

where $\hat{\mathbf{x}}_i$, $\overline{\boldsymbol{\sigma}}_i^m$ represent the integrated joint representations and the average variance of the multivariate Gaussian distribution of m modality. In this way, we preliminarily integrate stable unimodal representations into joint representations, where the contributions of different modalities are estimated dynamically.

Joint Representation Compression. In the SUFA module, we fully consider the consistency of semantics in different modalities to avoid noise in unimodal data. However,

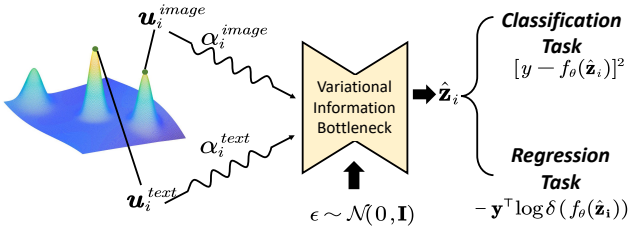

Figure 4. An illustration of the RMFI module. We provide Cross-Entropy loss and L_1 loss for classification-based and regression-based downstream tasks respectively.

with similar multivariate Gaussian distributions among different modalities, redundant duplicated information will be introduced into joint representations. Hence, we devise a joint representation compression with the Variational Information Bottleneck (VIB) [20]. Specifically, given the preliminary joint representation $\hat{\mathbf{x}}_i$ and the target observations $\mathbf{y}$ in label space, we learn the compressed joint representation $\hat{\mathbf{z}}$ in a latent space:

$$p(\hat{\mathbf{z}}_i|\hat{\mathbf{x}}_i) \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\sigma}}_i^2\mathbf{I}), \quad \hat{\boldsymbol{\mu}}_i = f_{\hat{\theta}_1}(\hat{\mathbf{x}}_i), \quad \hat{\boldsymbol{\sigma}}_i = f_{\hat{\theta}_2}(\hat{\mathbf{x}}_i). \tag{7}$$

where $f_{\hat{\theta}_1}$ and $f_{\hat{\theta}_2}$ are two fully connected layers. Similarly, we also adopt the re-parameterization tricks [40] for the final compressed joint representations:

$$\hat{\mathbf{z}}_i = \hat{\boldsymbol{\mu}}_i + \epsilon\hat{\boldsymbol{\sigma}}_i, \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{8}$$

As our proposed method can be applied to different downstream tasks, here we provide two training objectives for the joint representation compression. Specifically, we employ the cross-entropy for the classification-based task:

$$L_{\text{JRC}} = -\mathbf{y}^\top \log \delta(f_\theta(\hat{\mathbf{z}}_i)) + \lambda KL(p(\hat{\mathbf{z}}_i|\hat{\mathbf{x}}_i)\|\mathcal{N}(0, \mathbf{I})), \tag{9}$$

where $\delta(\cdot)$ represents softmax function and $f_\theta(\cdot)$ is the deep learning models for classification, λ is a hyperparameter. As for the regression-based task, we employ the mean square error as the training objective:

$$L_{\text{JRC}} = [y - f_\theta(\hat{\mathbf{z}}_i)]^2 + KL(p(\hat{\mathbf{z}}_i|\hat{\mathbf{x}}_i)\|\mathcal{N}(0, \mathbf{I})). \tag{10}$$

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate our method on five multimodal datasets, including two tri-modal datasets, *i.e.*, CMU-MOSI [41] and CMU-MOSEI [42] and three bi-modal datasets, *i.e.*, MVSA-Single [43], UPMC Food101 [1], and NYU Depth v2 [44] datasets. The *CMU-MOSI* [41] and *CMU-MOSEI* [42] datasets are widely employed video datasets for Multimodal Sentiment Analysis (MSA) task. The former encompasses 2199 short videos, each accompanied by a sentiment strength score ranging from -3 to 3, offering a nuanced measure of emotional intensity. The latter, designed specifically for emotion analysis tasks, is a larger dataset comprising 22,856 movie review clips. The bi-modal dataset, *i.e.*, *MVSA-Single* [43] is also used for MSA task but it only contains image-text pairs obtained from social media for sentiment classification. Following [2–5], we also employ the UPMC Food101 [1] and the NYU Depth v2 [44] datasets for Multimodal Image Classification (MIC) task. The *UPMC Food101* dataset contain 101 categories , respectively, these datasets are specifically designed for bi-modal classification tasks. The *UPMC FOOD101* dataset contains images with 101 categories obtained by Google Image Search and their corresponding textual descriptions. The NYU Depth v2 dataset [44] is utilized for scene recognition and comprises depth and RGB images. Following [5], we also adopt the commonly used 9 out of the 27 scene categories and the remaining categories as "Others".

Implementation Details. In alignment with alternative methodologies, for our trimodal video dataset, we employ FACET, COVAREP, and BERT as feature extractors for visual, audio, and text modalities, respectively. For the bi-modal dataset, we utilize ResNet-152 as the feature extractor for RGB and depth images, coupled with BERT as the feature extractor for text. As for the hyperparameters, we set the temperature factor τ and the balance factor λ to 0.5 and 1e-3 respectively. We employed the Adam optimizer with learning rate of 1e-5,and adopt the Reduce-on-Plateau learning rate adjustment strategy to train our EAU method.

Evaluation Metrics. For the regression-based MSA task based on CMU-MOSI and CMU-MOSEI, we follow the previous work [6, 7, 9, 45] and adopt Acc7, F1 score, and Pearson correlation coefficient as our metrics. As for the classification-based MSA task on the MSVA-Single dataset and MIC task on the UPMC FOOD101 and The NYU Depth v2 datasets, we reported commonly used metrics, including accuracy and F1 score.

4.2. Comparisons with State-Of-The-Arts

To show the superiority of our proposed EAU method on multimodal fusion, we compared our EAU method with existing state-of-the-art methods of the MSA task on the CMU-MOSI and CMU-MOSEI datasets, including MIB [28], HMA [6], MIM [7], GCNet [46], ConFEDE [8], DiCMoR [45], and DMD [9]. Moreover, we also make fair comparisons with the widely adopted simple multimodal fusion strategies *i.e.*, Concat and Late Fusion, as well as recent well-designed multimodal fusion strategies on the bi-modal datasets, including TMC [3], ITIN [47], MMBT [2], PMF [4], MVCN [48], and QMF [5]. Particularly, similar to the counterpart method QMF [5], we evaluate our method on noisy datasets to observe the model robustness.

Multimodal Fusion on Classification Task. We report the performance of classification tasks, including both MSA

Method	CMU-MOSI			CMU-MOSEI		
	Acc7	F1	Corr	Acc7	F1	Corr
MIB [28] (2022)	48.6	85.3	0.798	54.1	86.2	0.790
HMA [6] (2023)	45.3	85.6	0.782	52.8	85.4	0.787
MIM [7] (2023)	47.0	85.9	0.805	52.5	86.3	0.792
GCNet [46] (2023)	44.9	85.1	-	51.5	85.2	-
ConFEDE [8] (2023)	42.3	85.5	0.784	**54.9**	85.8	0.780
DiCMoR [45] (2023)	45.3	85.6	-	53.4	85.1	-
DMD [9] (2023)	45.6	86.0	-	54.5	86.6	-
EAU (Ours)	**48.8**	**86.2**	**0.809**	54.8	**86.9**	**0.816**

Method	MVSA-Single		Food 101		NYU Depth v2	
	Acc	F1	Acc	F1	Acc	F1
Concat	65.59	65.43	88.20	88.19	70.30	69.82
Late Fusion	76.88	75.72	90.69	90.77	69.14	68.32
MMBT [2] (2020)	78.50	-	91.52	91.28	71.04	-
TMC [3] (2021)	76.06	74.55	89.86	89.80	71.06	69.83
ITIN [47] (2022)	75.19	74.97	-	-	-	-
PMF [4] (2023)	-	-	91.68	-	-	-
MVCN [48] (2023)	76.06	74.55	-	-	-	-
QMF [5] (2023)	78.07	77.18	92.92	92.93	70.09	68.65
EAU (Ours)	**79.15**	**78.36**	**93.20**	**93.18**	**72.05**	**70.63**

Table 1. Comparisons with state-of-the-art multimodal fusion methods on MSA and MIC tasks. The CMU-MOSI and CMU-MOSEI datasets contain videos, audios, and texts. The MVSA-Single and Food-101 datasets consist of texts and RGB images. The NYU Depth v2 contains RGB and depth images. Note that CMU-MOSI and CMU-MOSEI are used for regression-based MSA tasks, while the others are used for classification-based MSA or MIC tasks.

and MIC tasks on bi-modal datasets in Table 1, where the best results are marked in bold. From the experimental results, we can observe that our proposed EAU method outperforms all three datasets. Particularly, compared with the recent state-of-the-art method QMF method [5], our EAU method achieves at least 1% absolute improvements on the MVSA-Single and NYU Depth v2 datasets. These results demonstrate that our EAU approach conducts better multimodal fusion on text, depth, and RGB images, affirming the effectiveness of leveraging aleatoric uncertainty in multimodal data. However, we also note that the improvements on UPMC Food 101 are not as remarkable as the other two benchmark datasets. We speculate that one probable reason is the images and texts in this dataset are clear compared with the others. Hence, our EAU method obtains fewer improvements since the augmentation is limited.

Multimodal Fusion on Regression Task. We evaluate our method on the regression task on the CMU-MOSEI and CMU-MOSI datasets, where the MSA task is conducted by predicting sentiment strengths. From the experimental results in Table 1, we can observe that: For the CMU-MOSI dataset, our proposed EAU method achieves state-of-the-art performance across all evaluation metrics, with a particularly significant improvement noted in Acc7. On the CMU-MOSEI dataset, our method has also demonstrated substantial improvements on most evaluation metrics. These results demonstrate the superiority of our proposed EAU, which benefits from high-quality multimodal fusion via learning stable and robust joint representations. Simultaneously, these results indicate that our approach can effortlessly extend to other modalities and can be seamlessly transferred to frameworks accommodating kinds of modalities.

Model Robustness on Noisy Multimodal Datasets. To validate the effectiveness of our model in handling data noise, we conducted more evaluation on the noisy datasets following a recent multimodal fusion strategy QMF [5]. Concretely, we consider different intensities of Gaussian and Salt-Pepper noise on the MVSA-Single and NYU Depth v2 datasets. For fair comparisons, we conducted 10

Noisy MVSA-Single					
Method	Clean	Salt-Pepper Noise		Gaussian Noise	
	$\epsilon = 0$	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = 5$	$\epsilon = 10$
Concat	65.59	58.69	51.16	50.70	46.12
Late Fusion	76.88	67.88	55.43	63.46	55.16
MMBT [2] (2020)	78.50	74.07	51.26	71.99	55.34
TMC [3] (2021)	74.87	68.02	56.62	66.72	60.35
QMF [5] (2023)	78.07	73.90	60.41	73.85	61.28
EAU (Ours)	**79.15**	**74.81**	**61.04**	**73.89**	**62.04**

Noisy NYU Depth v2					
Method	Clean	Salt-Pepper Noise		Gaussian Noise	
	$\epsilon = 0$	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = 5$	$\epsilon = 10$
Concat	70.44	57.98	44.51	59.97	53.20
Late fusion	69.16	56.27	41.22	59.63	51.99
MMTM [2] (2020)	71.04	59.45	44.59	60.37	52.28
TMC [3] (2021)	71.06	59.34	44.65	61.04	53.36
QMF [5] (2023)	70.09	58.50	45.69	61.62	55.60
EAU (Ours)	**72.05**	**59.83**	**46.85**	**63.33**	**58.85**

Table 2. Comparisons with state-of-the-arts concerning model performance on noisy MVSA-Single and NYU Depth v2 datasets.

experiments with different random seeds and reported the mean results as our final performance. According to the experimental results listed in Table 2, we can observe that our method has achieved state-of-the-art performance under various types and intensities of noise. Moreover, with the intensity of noise increasing, the margin of performance between our method and existing multimodal fusion strategies goes larger. Particularly, our method achieves more than 3% improvement on the Noisy NYU Depth v2 dataset over the QMF method [5]. Such a result demonstrates our method reveals better robustness against data noise. The reason is our model pioneeringly leverages the intrinsic aleatoric uncertainty in the training data to enhance the stable feature representation, which is not fully considered in existing counterpart methods [2, 3, 5].

4.3. Further Analysis

Analysis on Model Structure. To explore the impact of different model structures, we decomposed our proposed method into two components, SUFA and RMFI, and evalu-

Method	Clean	Salt-Pepper Noise		Gaussian Noise	
	Acc@$\epsilon = 0$	Acc@$\epsilon = 5$	Acc@$\epsilon = 10$	Acc@$\epsilon = 5$	Acc@$\epsilon = 10$
Naive Backbone + Late Fusion	76.88 ± 1.30	67.88 ± 1.87	55.43 ± 1.94	63.46 ± 3.46	55.16 ± 3.60
Naive Backbone + Concat	74.53 ± 0.97	62.24 ± 2.47	53.14 ± 2.89	56.08 ± 1.89	49.21 ± 2.81
Naive Backbone + RMFI	77.23 ± 0.74	70.99 ± 2.77	59.11 ± 1.92	66.53 ± 2.26	52.02 ± 3.45
SUFA + Late Fusion	77.11 ± 1.05	70.05 ± 1.50	55.93 ± 2.25	62.27 ± 4.20	55.90 ± 3.67
SUFA + Concat	78.18 ± 1.08	73.10 ± 1.07	60.67 ± 1.38	71.40 ± 1.48	59.11 ± 1.60
BERT (NAACL'19) [49]	75.61 ± 0.53	69.50 ± 1.50	47.41 ± 0.79	69.50 ± 1.50	47.41 ± 0.79
MMBT (arXiv'19) [2]	78.50 ± 0.40	74.07 ± 1.12	51.26 ± 5.65	71.99 ± 1.04	55.34 ± 2.84
TMC (ICLR'21) [3]	74.87 ± 2.24	68.02 ± 3.07	56.62 ± 3.67	66.72 ± 4.55	60.35 ± 2.79
QMF (ICML'23) [5]	78.07 ± 1.10	73.90 ± 1.89	60.41 ± 2.63	73.85 ± 1.42	61.28 ± 2.12
SUFA + RMFI (EAU)	**79.15 ± 0.60**	**74.81 ± 1.59**	**61.04 ± 1.31**	**73.89 ± 0.96**	**62.04 ± 1.26**

Modality	Acc	F1	Corr
T	44.1	83.7	0.785
V	16.7	45.3	0.072
A	15.8	48.7	0.099
T+A	46.7	84.2	0.794
T+V	46.3	84.6	0.786
A+V	20.2	55.3	0.117
T+A+V	**48.8**	**86.2**	**0.809**
T	76.30	75.90	-
V	63.58	63.35	-
T+V	**79.15**	**78.36**	-

Table 3. Analysis concerning key model structures in our EAU method on the Noisy MSVA-Single (Left), and analysis concerning modalities on the CMU-MOSI (Right Upper) and MVSA-Single (Right Lower) datasets. Note that BERT [49] adopts text modality only.

ated their effectiveness in combination with various model structures on the Noisy MVSA-Single dataset. Specifically, we employ two alternative designs for SUFA and RMFI respectively: *(1) Naive Backbone*: deploying the feature extractors directly, where the distributional representation and stable feature augmentation in the SUFA module are disabled. *(2) Late Fusion*: the unimodal representations are used for downstream tasks first and the fusion operations are conducted in probability space. To observe the stability of our method, we conduct 10 experiments with different random seeds and report the mean and variance in Table 3. Here we also make fair comparisons with several counterpart methods [2, 3, 5, 49] to furtherly show our superiority.

According to the experimental results presented in Table 3, we can observe that: (1) Our proposed SUFA and RMFI modules consistently improve performance on all metrics. Particularly, with the SUFA module, the results of fusion strategy *Concat* are boosted with more than 10% absolute performance on the noisy data. Moreover, compared with *Late Fusion*, our proposed RMFI module can also significantly boost the classification accuracy on noisy data with the SUFA module. These results prove again that our method shows great superiority in the effectiveness and robustness of multimodal fusion. (2) It also explicates that our proposed EAU method remarkably outperforms compared with the counterpart methods [2, 3, 5]. Particularly, compared with the recent state-of-the-art method QMF [5], our complete model reveals better performance and stability, demonstrating the capability of our EAU method again.

Analysis on the Effectiveness of Multimodal Fusion. To verify the effectiveness of multimodal fusion in our EAU method, we also conduct ablation studies concerning different modalities on the CMU-MOSI datasets, which consist of video (V), text (T), and audio (A) three modalities. By observing the experimental results illustrated in Table 3, we have drawn the following conclusions: (1) Our fusion strategy demonstrated significant effectiveness across different modal combinations. As the number of modalities increased, our approach facilitated the integration of multimodal information, resulting in a significant improvement across all evaluation metrics. (2) We also note that the text modality performs an essential role in the MSA task on the CMU-MOSI dataset. However, combined with the other two modalities via our proposed EAU method, the results are also boosted significantly, demonstrating the superiority of our EAU method again. Such a phenomenon also proves the rationality of our Dynamic Multimodal Integration process in the RMFI module.

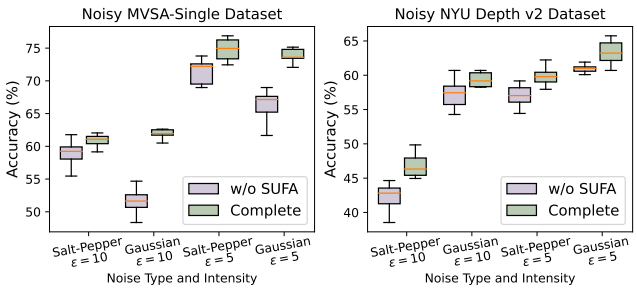

Figure 5. Analysis for robustness evaluated by over 10 times random experiments under different noisy data.

Analysis on Model Robustness. To verify the robustness and stability of our proposed EAU method, we conduct more ablation studies on the MVSA-Single and NYU Depth v2 datasets under different noise levels and observe the sparsity of final accuracy. Specifically, we conduct more than 10 experiments randomly with complete EAU (denoted as *Complete*) and EAU w/o SUFA (denoted as *w/o SUFA*), and show the statistics in Fig. 5. By observing the experimental results, we can find that: (1) The SUFA module consistently shows its effectiveness on both two datasets under different noise strengths. Particularly, when introducing more noise to the data, the ablated model shows strong fluctuations. Contrastively, the complete EAU model reveals remarkably reduced fluctuations, which demonstrates the superiority of model robustness and stability. (2) We also note that the improvements on the NYU Depth v2 are slightly lower than the MVSA-Single dataset, though both two datasets have significant results. We speculate this is caused by the difference between text modality and depth images. Concretely, the depth images can be regarded as another view of spatial information, which has less complementary information for RGB images compared with texts.

To this end, the introduced noise will show more impact on the depth-RGB than the image-text compositions.

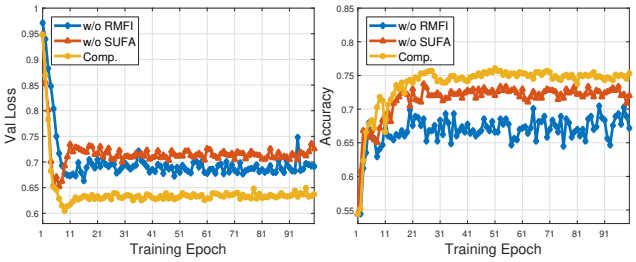

Figure 6. Analysis concerning the training convergence and performance fluctuations during training.

Analysis on Training Process. We also illustrate the training process of our proposed EAU method on the MVSA-Single in Fig 6 to observe the training convergence and performance fluctuations. According to the illustrated experimental results, we can observe that the complete EAU model exhibits a smoother training process and converges more rapidly with a remarkably better performance. Particularly, compared with the ablated model *w/o RMFI*, the other two models show much better stability during training. This is because the fused features learned by the ablated model *w/o RMFI* have redundant information caused by cross-modal distributional alignment in the SUFA module. It proves again the rationality of our proposed RMFI module, which deploys an information-theoretic strategy to address the limitation.

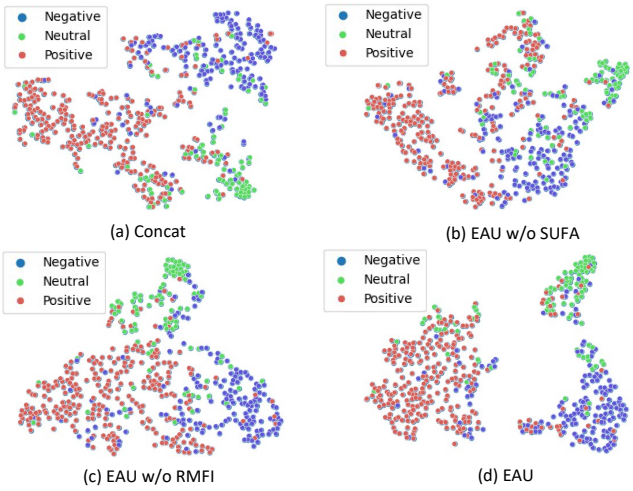

Figure 7. Visualizations of joint representations by t-SNE [50] on MVSA-Single dataset. Here we adopt *Concat* for the ablated model *w/o RMFI*.

Visualizations of Joint Representations. To further verify the superiority of our proposed EAU method, we also employ the t-SNE [50] to visualize the learned joint representations. As illustrated in Fig. 7, it can be observed that the features generated by our method exhibit a more compact and discriminative distribution. In contrast, when the SUFA or RMFI module is removed, more sparse points appear around the feature distribution which leads to corrupted classification results at last. Such results indicate that our method is more capable of learning a representative joint representation.

Figure 8. Visualizations of test cases selected from the MSVA-Single datasets. It can be observed that our EAU method reveals better robustness against the noise.

Qualitative Analysis. Moreover, we also illustrate several typical test cases in Fig. 8 and make comparisons with the recent counterpart method QMF [5]. Under different noise settings, our proposed EAU method consistently classifies the sentiment of the given image-text input precisely, while the QMF method returns incorrect results. It explicates that our proposed EAU method benefits from the well-designed SUFA and RMFI module, which is effective in learning more stable and robust joint representations for noise-resistant performance with higher accuracy.

5. Conclusion

In this paper, we proposed a novel multimodal fusion method, namely, Embracing Aleatoric Uncertainty (EAU). It achieves more discriminative joint representations by fully considering the aleatoric uncertainty in multimodal data. Specifically, with the well-designed Stable Unimodal Feature Augmentation (SUFA) and Robust Multimodal Feature Integration (RMFI) modules, our proposed EAU could learn compact and robust joint representations. We evaluated our proposed EAU method on five multimodal benchmark datasets for both classification and regression tasks and demonstrated its superiority in fusion performance and robustness. For future work, we will further explore the uncertainties in multimodal learning.

6. Acknowledgement

This work was sponsored in part by the National Natural Science Foundation of China under Grants (No. 62222203 and No. 62072080) and the New Cornerstone Science Foundation through the XPLORER PRIZE.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461, 2014. 1, 5

[2] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019. 1, 2, 5, 6, 7

[3] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. *arXiv preprint arXiv:2102.02051*, 2021. 1, 2, 5, 6, 7

[4] Yaowei Li, Ruijie Quan, Linchao Zhu, and Yi Yang. Efficient multimodal fusion via interactive prompting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2023. 1, 5, 6

[5] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. *arXiv preprint arXiv:2306.02050*, 2023. 1, 2, 4, 5, 6, 7, 8

[6] Ronghao Lin and Haifeng Hu. Dynamically shifting multimodal representations via hybrid-modal attention for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 2023. 1, 5, 6

[7] Ying Zeng, Sijie Mai, Wenjun Yan, and Haifeng Hu. Multimodal reaction: Information modulation for cross-modal representation learning. *IEEE Transactions on Multimedia*, 2023. 1, 5, 6

[8] Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 7617–7630, 2023. 1, 5, 6

[9] Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6631–6640, 2023. 1, 5, 6

[10] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020. 1, 2

[11] 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 Conference on Computer Vision and Pattern Recognition*, pages 8238–8247, 2022. 1, 2

[12] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International Conference on Machine Learning*, pages 9226–9259. PMLR, 2022. 1, 2

[13] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. PMR: prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 20029–20038. IEEE, 2023. 1, 2

[14] 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. 2, 3

[15] Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. Data uncertainty learning in face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5710–5719, 2020. 2

[16] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6902–6911, 2019. 2

[17] Yatai Ji, Junjie Wang, Yuan Gong, Lin Zhang, Yanru Zhu, Hongfa Wang, Jiaxing Zhang, Tetsuya Sakai, and Yujiu Yang. Map: Multimodal uncertainty-aware vision-language pre-training model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 23262–23271, 2023. 2

[18] Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. *arXiv preprint arXiv:1412.6623*, 2014. 2

[19] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 2, 3

[20] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016. 2, 3, 5

[21] Xun Jiang, Xing Xu, Zhiguo Chen, Jingran Zhang, Jingkuan Song, Fumin Shen, Huimin Lu, and Heng Tao Shen. DHHN: dual hierarchical hybrid network for weakly-supervised audio-visual video parsing. In *ACM International Conference on Multimedia*, pages 719–727, 2022. 2

[22] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Collecting cross-modal presence-absence evidence for weakly-supervised audio-visual event perception. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 18827–18836, 2023. 2

[23] Zhiguo Chen, Xun Jiang, Xing Xu, Zuo Cao, Yijun Mo, and Heng Tao Shen. Joint searching and grounding: Multi-granularity video content retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 975–983, 2023. 2

[24] Shenshen Li, Xing Xu, Xun Jiang, Fumin Shen, Xin Liu, and Heng Tao Shen. Multi-grained attention network with mutual exclusion for composed query-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[25] Shenshen Li, Xing Xu, Xun Jiang, Fumin Shen, Zhe Sun, and Andrzej Cichocki. Cross-modal attention preservation with self-contrastive learning for composed query-based image retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(6), 2024. 2

[26] Xing Xu, Kaiyi Lin, Yang Yang, Alan Hanjalic, and Heng Tao Shen. Joint feature synthesis and embedding: Adversarial cross-modal retrieval revisited. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3030–3047, 2022. 2

[27] Shenshen Li, Xing Xu, Xun Jiang, Fumin Shen, Xin Liu, and Heng Tao Shen. Multi-grained attention network with mutual exclusion for composed query-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023. 2

[28] Sijie Mai, Ying Zeng, and Haifeng Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 2022. 2, 3, 5, 6

[29] Xun Jiang, Xing Xu, Jingran Zhang, Fumin Shen, Zuo Cao, and Heng Tao Shen. Sdn: Semantic decoupling network for temporal language grounding. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2022. 2

[30] Xun Jiang, Xing Xu, Jingran Zhang, Fumin Shen, Zuo Cao, and Heng Tao Shen. Semi-supervised video paragraph grounding with contrastive encoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2456–2465. IEEE, 2022. 2

[31] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems*, pages 14200–14213, 2021. 2, 3

[32] Guanglei Yang, Enrico Fini, Dan Xu, Paolo Rota, Mingli Ding, Moin Nabi, Xavier Alameda-Pineda, and Elisa Ricci. Uncertainty-aware contrastive distillation for incremental semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2567–2581, 2022. 2

[33] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. Dual-evidential learning for weakly-supervised temporal action localization. In *European Conference on Computer Vision*, pages 192–208. Springer, 2022. 2

[34] Yifei Huang, Lijin Yang, and Yoichi Sato. Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 18908–18918, 2023. 2

[35] Shenshen Li, Xing Xu, Yang Yang, Fumin Shen, Yijun Mo, Yujie Li, and Heng Tao Shen. DCEL: deep cross-modal evidential learning for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6292–6300, 2023. 2

[36] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016. 3

[37] Yuchen Shen, Dong Zhang, Zhihao Song, Xuesong Jiang, and Qiaolin Ye. Learning to reduce information bottleneck for object detection in aerial images. *IEEE Geoscience and Remote Sensing Letters*, 2023. 3

[38] Aming Wu and Cheng Deng. Tib: Detecting unknown objects via two-stream information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[39] Sungtae An, Nataraj Jammalamadaka, and Eunji Chong. Maximum entropy information bottleneck for uncertainty-aware stochastic embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3808–3817, 2023. 3

[40] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4, 5

[41] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016. 5

[42] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2236–2246, 2018. 5

[43] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. Sentiment analysis on multi-view social data. In *International Conference on Multimedia Modeling*, pages 15–27, 2016. 5

[44] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760, 2012. 5

[45] Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 22025–22034, 2023. 5, 6

[46] Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 5, 6

[47] Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, Hantao Liu, and Jiansheng Qian. Multimodal sentiment analysis with image-text interaction network. *IEEE Transactions on Multimedia*, 2022. 5, 6

[48] Yiwei Wei, Shaozu Yuan, Ruosong Yang, Lei Shen, Zhangmeizhi Li, Longbiao Wang, and Meng Chen. Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 5240–5252, 2023. 5, 6

[49] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 7

[50] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8