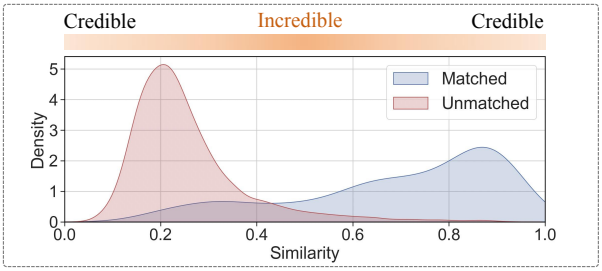
Fuzzy Multimodal Learning for Trusted Cross-modal Retrieval

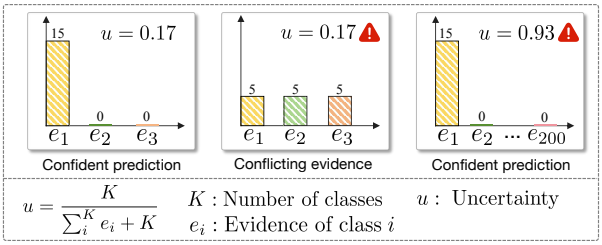
Siyuan Duan[1] Yuan Sun[1] Dezhong Peng[1,2,3] Zheng Liu[2] Xiaomin Song[2] Peng Hu[1*]
[1]College of Computer Science, Sichuan University, Chengdu, 610095, China.
[2]Sichuan National Innovation New Vision UHD Video Technology Co., Ltd, Chengdu, 610041, China
[3]Tianfu Jincheng Laboratory, Chengdu, 610093, China.

Abstract

Cross-modal retrieval aims to match related samples across distinct modalities, facilitating the retrieval and discovery of heterogeneous information. Although existing methods show promising performance, most are deterministic models and are unable to capture the uncertainty inherent in the retrieval outputs, leading to potentially unreliable results. To address this issue, we propose a novel framework called FUzzy Multimodal lEarning (FUME), which is able to self-estimate epistemic uncertainty, thereby embracing trusted cross-modal retrieval. Specifically, our FUME leverages the Fuzzy Set Theory to view the outputs of the classification network as a set of membership degrees and quantify category credibility by incorporating both possibility and necessity measures. However, directly optimizing the category credibility could mislead the model by over-optimizing the necessity for unmatched categories. To overcome this challenge, we present a novel fuzzy multimodal learning strategy, which utilizes label information to guide necessity optimization in the right direction, thereby indirectly optimizing category credibility and achieving accurate decision uncertainty quantification. Furthermore, we design an uncertainty merging scheme that accounts for decision uncertainties, thus further refining uncertainty estimates and boosting the trustworthiness of retrieval results. Extensive experiments on five benchmark datasets demonstrate that FUME remarkably improves both retrieval performance and reliability, offering a prospective solution for cross-modal retrieval in high-stakes applications. Code is available at https://github.com/siyuancncd/FUME .



(a) Problem of incredible results in cross-modal retrieval.



(b) Counter-intuitive problem of Evidential Deep Learning in uncertainty estimation.

Figure 1. (a) Illustration of the problem of incredible results in cross-modal retrieval. Experiments are conducted by GNN4CMR [5] on the Pascal Sentence dataset [10]. This figure reveals that relying solely on inter-modality similarity scores in the common space can cause mismatches for challenging samples, a problem that most methods fail to address, leading to incredible retrieved results. (b) Illustration of the counter-intuitive problem of EDL [11] for uncertainty estimation. In the middle case, conflicting evidence should intuitively result in high uncertainty; however, the calculated uncertainty is low. Conversely, in the right case, confident prediction should intuitively yield low uncertainty, yet the estimated uncertainty is high.

1. Introduction

Cross-modal retrieval (CMR) [1] seeks to search for related samples across different modalities, which has garnered significant attention from both academia and industry, driven by the exponential growth of multimedia data, including images, text, and audio [2, 3]. The primary challenge in CMR lies in calculating the similarity between heterogeneous samples [4]. To this end, most existing methods [5–9] address this by projecting different modalities into a shared latent space, enabling semantically related heterogeneous samples to be represented with similar embeddings.

Although these methods have shown promising performance, they are deterministic models, relying solely on sim-

*Corresponding author: Peng Hu (penghu.ml@gmail.com).

ilarity scores while neglecting the possibility of unreliable or untrustworthy outcomes. This limitation is especially concerning in applications where reliability is crucial, such as healthcare, autonomous driving, and other high-stakes, cost-sensitive applications. However, as shown in Figure 1 (a), the existing similarity-based methods cannot discern challenging samples due to the epistemic uncertainty inherent in the models, leading to unreliable outcomes. To address this, an intuitive solution is to quantify the epistemic uncertainty associated with retrieved results.

To quantify uncertainty, various methods have been developed using different techniques [12], including Bayesian deep learning [13], MC-dropout [14] and Deep Ensembles [15]. While these approaches are effective in uncertainty estimation, they often suffer from high computational costs. To address this issue, Evidential Deep Learning (EDL) [11] offers an alternative by treating neural network predictions as subjective opinions and directly inferring uncertainty. However, EDL's uncertainty estimation depends solely on total evidence and the number of classes. This implies that the uncertainty is underestimated when the number of categories is low relative to the total evidence and overestimated when the total evidence is low relative to the number of categories. As Figure 1 (b) shows, it's counter-intuitive that conflicting evidence would lead to low-level uncertainty (middle case), and confident prediction would result in high-level uncertainty (right case).

To address the problems mentioned above, we propose a novel framework to quantify cross-modal uncertainty based on the Fuzzy Set Theory [16], called FUzzy Multi-modal lEarning (FUME), enabling trusted cross-modal retrieval. The architecture of FUME is illustrated in Figure 2. Specifically, to capture decision uncertainty and resolve the counter-intuitive issues in EDL, we incorporate Fuzzy Set Theory to view the outputs of the classification network as a set of membership degrees and quantify category credibility by considering both possibility and necessity measures. However, directly optimizing category credibility could mislead the model by over-optimizing the necessity for unmatched categories. To address this, we propose a novel fuzzy multimodal learning strategy that leverages label information to guide necessity optimization in the right direction, thereby indirectly optimizing the category credibility. Moreover, we utilize the normalized entropy of the category credibility as a metric for decision uncertainty, which avoids the counter-intuitive issues inherent in EDL. Finally, we incorporate classifiers into the CMR model to quantify decision uncertainty during the retrieval process and design an uncertainty merging scheme that accounts for uncertainty across two modalities, providing an accurate cross-modal uncertainty quantification. This uncertainty, along with similarity scores, enables trusted and reliable cross-modal retrieval. In summary, the primary contributions of this work are as follows:

- We reveal and investigate a rarely studied yet practically significant issue—the occurrence of incredible results in CMR. To the best of our knowledge, our FUME could be one of the first methods that focus on quantifying the epistemic uncertainty associated with CMR results.
- To capture epistemic uncertainty, we propose category credibility along with a tailored loss function for optimization. Our FUME not only inherits the low computational cost and direct uncertainty inference but also avoids the counter-intuitive problem associated with EDL.
- We propose a novel metric, cross-modal uncertainty, which is leveraged to promote trusted and reliable retrieval by accurately capturing the uncertainty of each retrieved result.
- Extensive experiments demonstrate that FUME achieves superior precision and reliability, attributed to its robust performance and accurate uncertainty estimation.

2. Related Work

2.1. Cross-modal Retrieval

Cross-modal retrieval methods typically involve projecting different modalities into a shared latent space where their similarities can be measured [17–20]. Early solutions primarily employed linear transformations to map different modalities into common representations [21–23]. With the advent of deep learning, recent methods have leveraged its nonlinear learning capability to bridge the cross-modal gap, including unsupervised [24–26], semi-supervised [27–29], supervised [1, 2, 6, 18, 19, 30], along with a variety of hashing methods [31–34]. Among supervised methods, Deep Supervised Cross-Modal Retrieval (DSCMR) [1] takes advantage of label information for cross-modal retrieval but suffers from retraining requirements when encountering new modalities. To address this, Scalable Deep Multimodal Learning (SDML) [18] decouples modality-specific training for better scalability. However, these models are often sensitive to noise between modalities. The Deep Evidential Cross-modal Learning framework (DECL) [30] introduces the Dempster-Shafer Theory of Evidence to model the uncertainty in retrieved results, improving the robustness. Nevertheless, it lacks fine-grained uncertainty modeling for each individual result, which is crucial for trusted cross-modal retrieval.

2.2. Uncertainty-based Deep Learning

Recently, uncertainty-based deep learning has seen extensive research, with a range of techniques proposed over the years [12]. Bayesian deep learning offers a principled way to estimate uncertainty by applying a distribution over model parameters, but its high computational costs limit widespread adoption [13]. To mitigate this, Monte Carlo

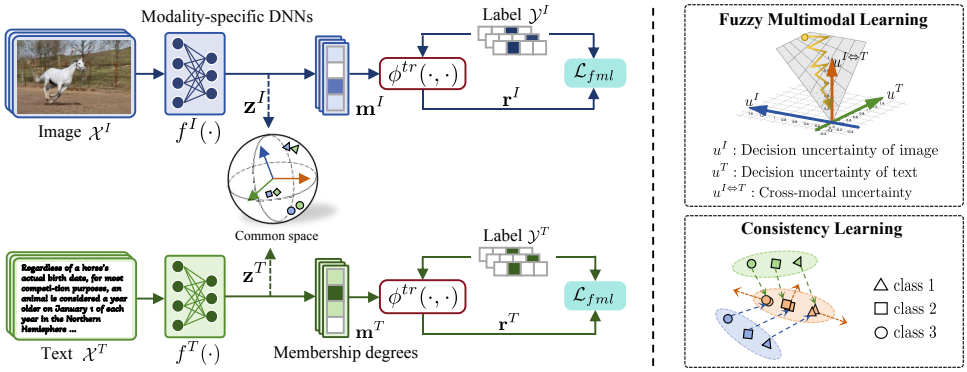


Figure 2. The overview of the proposed Fuzzy Multimodal learning (FUME) method. Firstly, modality-specific DNNs ($f^I(\cdot)$ and $f^T(\cdot)$) project image modality samples $\mathcal{X}^I$ and text modality samples $\mathcal{X}^T$ into a common space. Secondly, the representations ($\mathbf{z}^I$ and $\mathbf{z}^T$) in the common space map to membership degrees ($\mathbf{m}^I$ and $\mathbf{m}^T$). Finally, the membership degrees and labels ($\mathcal{Y}$) are inputted into the function $\phi^{tr}(\cdot,\cdot)$ to determine the category credibility during training ($\mathbf{r}^I$ and $\mathbf{r}^T$). These category credibilities are then used to compute loss $\mathcal{L}_{fml}$ and supervise the learning of the entire neural network. At the same time, Consistency learning is employed to eliminate the cross-modal discrepancy. During the optimization process, FUME will reduce the decision uncertainty of each modality, ultimately reducing the cross-modal uncertainty.

dropout (MC-dropout) introduces dropout layers to neural networks, using prediction distributions to estimate uncertainty [14]. However, without proper dropout layers, MC-Dropout can't effectively estimate uncertainty. Another popular method, Deep Ensembles [15], quantifies uncertainty by comparing predictions from multiple independently trained models. While effective, these methods are computationally intensive. In contrast, Evidential Deep Learning (EDL) [11], based on subjective logic theory, proposes an alternative approach by estimating uncertainty directly from the network outputs without the need for multiple models or weight sampling. Despite its success, EDL's reliance on total evidence and the number of classes often leads to counter-intuitive uncertainty estimation, making the predicted uncertainty inaccurate. Alternatively, Uncertainty Theory [35], which draws from Fuzzy Set Theory [16], provides a more nuanced perspective, incorporating both possibility and necessity measures, making it particularly well-suited for addressing complex scenarios in cross-modal retrieval.

3. The Proposed Method

3.1. Problem Formulation

For a clear presentation, we first define some notations as follows. Bold uppercase letters ($\mathbf{X}$) represent matrices, and bold lowercase letters ($\mathbf{x}$) denote column vectors. Consider a K-category multimodal dataset $D = \{\mathcal{M}_j\}_{j=1}^M = \{\mathcal{X}_j, \mathcal{Y}_j\}_{j=1}^M$, where M is the number of modalities and N is the number of samples in each modality. Specifically, $\mathcal{M}_j = \{(\mathbf{x}_i^j, \mathbf{y}_i^j)\}_{i=1}^N$ represents the set for the j-th modality, where $\mathbf{x}_i^j \in \mathbb{R}^{d_j}$ is the i-th sample from the j-th modality, d_j is the dimensionality of the j-th modality, and $\mathbf{y}_i^j = [y_{i1}^j, y_{i2}^j, ..., y_{iK}^j] \in \mathbb{R}^K$ is the semantic label vector of $\mathbf{x}_i^j$, with K being the number of categories. If the i-th sample belongs to the k-th category, y_{ik} = 1, otherwise $y_{ik} = 0$. The goal of CMR is to learn projection functions for each modality such that $\mathbf{z}^j = f^j(\mathbf{x}^j, \Theta^j) \in \mathbb{R}^L$, where $\mathbf{z}^j$ is the normalized representation in the common space, L is the dimensionality of the common space, and $\{\Theta^j\}_{j=1}^M$ are the trainable parameters for each modality. This enables the comparison of samples across different modalities in the shared space.

3.2. Fuzzy Multimodal Learning

3.2.1. Credibility Modeling

Fuzzy systems have the ability to effectively handle the uncertainties and ambiguities inherent in real-world data [36]. In the Fuzzy Set Theory [16], the membership degree quantifies how possible a sample belongs to a fuzzy set. Similarly, the output probabilities of a classification network, ranging from 0 to 1, represent the possibility of a sample belonging to each category, with higher values indicating a greater possibility of classification into that category. This similarity allows us to view the classifier's prediction for a category as a membership degree of a category. Given a sample $\mathbf{x}_i^j$, the membership degrees for each category can be represented as $m_{i1}^j, m_{i2}^j, ..., m_{iK}^j$, where K is the number of categories. However, membership degree only provides the possibility measure of an event (i.e., the possibility that the sample belongs to a certain category), rather than the necessity measure, which indicates the certainty that the sample does not belong to other categories [35]. To address this, we define the necessity belonging to each category as:

$$e_{ik}^j = 1 - \max\{m_{il}^j \mid l \neq k\},\ k = 1, ..., K, \qquad (1)$$

where $\max\{m_{il}^j \mid l \neq k\}$ is the highest membership degree among the other categories $\{l\}_{l\neq k}$. By integrating both possibility measure and necessity measure, we define the category credibility as:

$$c_{ik}^j = \frac{1}{2}\left(m_{ik}^j + 1 - \max\{m_{il}^j \mid l \neq k\}\right),\ k = 1, 2, ..., K, \tag{2}$$

which can be arranged as $\mathbf{c}_i^j = [c_{i1}^j, c_{i2}^j, ..., c_{iK}^j] \in \mathbb{R}^K$.

3.2.2. Credibility Learning

To map the representations in common space to the corresponding category, first, the representations are multiplied by an orthogonal weight matrix $\mathbf{W} \in \mathbb{R}^{K\times L}$. Subsequently, an activation function (i.e., ReLU) is applied to yield output values in the range of [0,1]. These values denote the membership degrees. The calculation formula is as follows :

$$\mathbf{m}_i^j = \text{ReLU}(\mathbf{W}\mathbf{z}_i^j), \tag{3}$$

where $\mathbf{z}_i^j = f(\mathbf{x}_i^j)$, $\mathbf{z}_i^j$ is the normalized representation of the i-th sample of j-th modality in the common space. The corresponding category credibility could be derived by Equation (2).

To learn discriminative representations for cross-modal retrieval, we aim for each sample to have the highest possible category credibility for the category to which it belongs while maintaining the lowest possible category credibility for the category to which it does not belong. Intuitively, this could be achieved by directly aligning the category credibility $\mathbf{c}_i^j$ with the corresponding one-hot labels $\mathbf{y}_i^j$, i.e., minimizing $||\mathbf{c}_i^j - \mathbf{y}_i^j||_2$. However, this approach risks over-optimizing the necessity for unmatched categories. Specifically, when $y_{ik}^j = 0$, m_{ik}^j would tend to 0, and the necessity $e_{ik}^j = 1 - \max\{m_{il}^j \mid l \neq k\}$ would also approach 0, forcing m_{il}^j to approach 1. This is problematic because m_{il}^j should approach 0 when $y_{il}^j = 0$, rather than 1, leading to suboptimal performance. To address this issue, we propose a fuzzy multimodal learning loss that optimizes the category credibilities while guiding the model toward the correct solution:

$$\mathcal{L}_{fml} = \frac{1}{N_b}\sum_{j}^{M}\sum_{i}^{N_b} ||\mathbf{r}_i^j - \mathbf{y}_i^j||_2, \tag{4}$$

where M denotes the number of modalities, N_b is the batch size, and $\mathbf{r}_i^j = \phi^{tr}(\mathbf{m}_i^j, \mathbf{y}_i^j) = [r_{i1}^j, r_{i2}^j, ..., r_{iK}^j] \in \mathbb{R}^K$ represents the category credibility during training, defined as:

$$r_{ik}^j = \begin{cases} \frac{m_{ik}^j + 1 - \max\{m_{il}^j | l\neq k\}}{2}, & if\ y_{ik}^j = 1, \\ \frac{m_{ik}^j + 1 - m_{il}^j}{2}, & if\ y_{ik}^j = 0, l = \arg\max\limits_k\ y_{ik}^j, \end{cases} \tag{5}$$

where $k = 1, 2, ..., K$. From Equation (4), one could see that this loss function could ensure m_{ik}^j approaches 0 for unmatched categories for $y_{ik}^j = 0$ and approaches 1 for the matched category $y_{ik}^j = 1$ by using label information. Specifically, when $y_{ik}^j = 0$ and $y_{il}^j = 1$ (where $l = \arg\max\limits_k\ y_{ik}^j$), we expect the membership degree of the matched category to be greater than that of any unmatched categories after training. For unmatched categories, the necessity should be calculated as $1 - m_{il}^j$, encouraging m_{il} to approach 1. In other word, m_{ik}^j should tend toward 1 when $y_{ik}^j = 1$, implying that $\max\{m_{il}^j \mid l \neq k\}$ should tend toward 0. Thus, this approach could direct the necessity optimization correctly, thereby avoiding mis-optimization of the model and achieving correct category credibility optimization.

Moreover, our method could reduce both decision uncertainty and cross-modal uncertainty by minimizing this loss function, as introduced in Section 3.5 and illustrated in the top right corner of Figure 2.

3.3. Consistency Learning

To eliminate the cross-modal discrepancy, we employ multimodal contrastive learning to project different modalities into a shared space by maximizing mutual information [19]. First, the probability of a sample $\mathbf{x}_i^j$ belonging to the j modality of the i-th instance is defined as :

$$P(i|\mathbf{x}_i^j) = \frac{\sum_{v=1}^{M} exp(\frac{1}{\tau}(\mathbf{z}_i^v)^T\mathbf{z}_i^j)}{\sum_{v=1}^{M}\sum_{t=1}^{N_b} exp(\frac{1}{\tau}(\mathbf{z}_t^v)^T\mathbf{z}_i^j)}, \tag{6}$$

where τ is a temperature parameter. Inspired by self-supervised learning methods [37, 38], we could align cross-modal samples from the same instance while pushing apart those from different instances to alleviate the cross-modal discrepancy and preserve instance-level discrimination in the common space. To achieve this, we minimize the following negative log-likelihood:

$$\mathcal{L}_{cl} = -\frac{1}{N_b}\sum_{j=1}^{M}\sum_{i=1}^{N_b} log\left(P(i|\mathbf{x}_i^j)\right), \tag{7}$$

thereby encouraging the model to compact positive (matched) cross-modal pairs while scattering negative (unmatched) pairs in the common space, as shown in Figure 2.

3.4. Optimization

By combining fuzzy multimodal learning and consistency learning, the final loss function could be formulated as:

$$\mathcal{L} = \mathcal{L}_{fml} + \alpha\mathcal{L}_{cl}, \tag{8}$$

where α is a positive hyperparameter to balance the contributions of $\mathcal{L}_{cl}$. Finally, our FUME could minimize Equation (8) to optimize the network parameters iteratively in a batch-by-batch manner by using gradient descent, as summarized in Algorithm 1.

Algorithm 1 Main optimization process of our FUME.

Input: The training multimodal data $\mathcal{D} = \{\mathcal{M}_j\}_{j=1}^{M}$, the dimensionality of representations in common space L, the orthogonal weight matrix $\mathbf{W} \in \mathbb{R}^{K\times L}$, batch size N_b, maximal epoch number N_e, temperature parameter τ, balance parameter α, learning rate η, and the modality-specific DNNs $\{f^j(\cdot,\Theta^j)\}_{j=1}^{M}$.

for $1, 2, \cdots, N_e$ **do**
- Randomly select N_b samples from every modality to construct a mini-batch $\{(\mathbf{x}_i^j, \mathbf{y}_i^j)_{i=1}^{N_b}\}_{j=1}^{M}$.
- Calculate the common representations $\{\{\mathbf{z}_i^j\}_{i=1}^{N_b}\}_{j=1}^{M}$ for all samples of the mini-batch by using their corresponding modality-specific DNNs $\{f^j(\cdot,\Theta^j)\}_{i=j}^{M}$, and normalize them.
- Calculate the membership degrees $\{\{\mathbf{m}_i^j\}_{i=1}^{N_b}\}_{j=1}^{M}$ by Equation (3).
- Calculate the category credibilities during training $\{\{\mathbf{r}_i^j\}_{i=1}^{N_b}\}_{j=1}^{M}$ by Equation (5).
- Compute $\mathcal{L}_{fml}$ and $\mathcal{L}_{cl}$ according to Equation (4) and Equation (7) on minibatch respectively.
- Update FUME parameters $\{\Theta^j\}_{j=1}^{M}$ using gradient descent algorithm with learning rate η.

end for

Output: Optimized network parameters $\{\Theta^j\}_{j=1}^{M}$.

3.5. Cross-modal Uncertainty Inference

In the inference stage, similarity is a widely used metric for traditional CMR methods to measure whether two samples from distinct modalities match, e.g., cosine similarity. However, these methods solely rely on similarity scores to retrieve the related information from the database while neglecting the uncertainty of the retrieved results, thus hindering trusted cross-modal retrieval. Intuitively, these methods could directly use the classification probability or similarity as the uncertainty. However, they cannot model the uncertainty in the predictions, leading to incorrect uncertainty estimation. On the contrary, our FUME could provide multi-dimensional predictions to measure the uncertainty inherent to results indirectly. In brief, we first quantify the decision uncertainty of prediction for each modality and subsequently fuse the decision uncertainty of two modalities to infer the cross-modal uncertainty of retrieved results.

Based on Shannon's entropy, which characterizes the uncertainty resulting from information deficiency, we define decision uncertainty. For convenience, we denote Shannon's funcion, $S(t) = -t\ln t-(1-t)\ln(1-t)$ with the convention that $0\cdot\ln 0 = 0$ [39]. It is evident that $t = 0.5$ serves as the symmetry axis of $S(t)$, and the maximum value of $\ln 2$ also occurs at $t = 0.5$. Using the function $S(t)$, we give the following definition of decision uncertainty:

Definition 1: Let $\mathbf{c}_i^j = [c_{i1}^j, c_{i2}^j, ..., c_{iK}^j] \in \mathbb{R}^K$ be the vector of category credibility of j-th modality and i-th sample and $\forall c_{ik}^j \in [0,1],\ k = 1, 2, ..., K$. Then, the decision uncertainty is defined by

$$\begin{aligned} u_i^j = U(\mathbf{c}_i^j) &= \frac{H(\mathbf{c}_i^j)}{K\cdot\ln 2} = \frac{\sum_{k=1}^{K} S(c_{ik}^j)}{K\cdot\ln 2} \\ &= \frac{\sum_{k=1}^{K} -c_{ik}^j\cdot\ln(c_{ik}^j) - (1-c_{ik}^j)\cdot\ln(1-c_{ik}^j)}{K\cdot\ln 2}, \end{aligned} \tag{9}$$

where $H(\mathbf{c}_i^j)$ is the entropy of category credibility and K is the number of categories. This uncertainty is in the range $[0, 1]$ [40].

The aforementioned decision uncertainty solely captures the uncertainty within a single modality, thus rendering it unsuitable for cross-modal retrieval applications. Therefore, a merging scheme is needed to consider the two modalities simultaneously. Let $u_i^1 \in [0,1]$ and $u_i^2 \in [0,1]$ represent the decision uncertainties for the i-th sample from two modalities, 1 and 2, respectively. The cross-modal uncertainty ($u_i^{1\Leftrightarrow 2}$) is expected to have the following reasonable properties:

- It ranges should from 0 to 1, i.e., $u_i^{1\Leftrightarrow 2} \in [0,1]$;
- It considers two different modalities, so it should be greater than or equal to either of the uncertainty of any single modal, i.e., $u_i^{1\Leftrightarrow 2} \geqslant \max\{u_i^1, u_i^2\}$;
- As long as one of the two modalities is definitely uncertain, the cross-modal uncertainty should be 1, i.e., if $u_i^1 = 1$ or $u_i^2 = 1$, $u_i^{1\Leftrightarrow 2} = 1$;
- When u_i^1 or u_i^2 increases, $u_i^{1\Leftrightarrow 2}$ also should increase, i.e., if $u_2^1 \geqslant u_1^1$ and $u_2^2 \geqslant u_1^2$, where $u_1^1,\ u_2^1,\ u_1^2$, and $u_2^2 \in [0,1]$, $u_2^{1\Leftrightarrow 2} \geqslant u_1^{1\Leftrightarrow 2}$.

Based on the above four properties, we propose a merge function $(g(\cdot,\cdot))$:

Definition 2: let $u_i^1 \in [0,1]$ and $u_i^2 \in [0,1]$ represent the decision uncertainties of the i-th sample from modalities 1 and 2, the cross-modal uncertainty across them is defined by

$$u_i^{1\Leftrightarrow 2} = g(u_i^1, u_i^2) = 1 - (1-u_i^1)(1-u_i^2). \tag{10}$$

Due to space limitations, the properties and proof of decision uncertainty and the proof of the above properties of cross-modal uncertainty are provided in the supplementary material.

4. Expriments

In this section, we validate the effectiveness of our FUME through comprehensive experiments. We evaluate our FUME on five widely used benchmark datasets and compare it against 13 state-of-the-art cross-modal retrieval methods. Without loss of generality, we focus on evaluating the accuracy and trustworthiness of retrieved results on two modalities: image and text.

4.1. Experiment Settings

4.1.1. Datasets

We conduct experiments on five widely-used datasets: **Pascal Sentence** [10], **Wikipedia** [41], **NUS-WIDE-10K** [42], **INRIA-Websearch** [43], and **XMediaNet** [44] to evaluate cross-modal retrieval performance. Additional dataset details are provided in the supplementary material.

4.1.2. Implementation Details

For feature extraction, we use VGGNet [45] for images and word2vec [46] for text. Each modality-specific subnet consists of four fully connected layers for the NUS-WIDE-10K [42] and INRIA-Websearch [43] datasets and three fully connected layers for the other three datasets. Each layer consists of 4,096 hidden units and employs ReLU [47] as the activation function. The temperature parameter (τ in Algorithm 1) is set as 1, while batch sizes of 100, 200, 300, 300, and 100 are applied to the Pascal Sentence [10], Wikipedia [41], NUS-WIDE-10K [42], INRIA-Websearch [43], and XMediaNet [44], respectively. The balance parameter (α in Algorithm 1) is set as 10, 0.05, 0.5, 0.05, and 1 for these datasets, respectively. The dimensionality of representations in the common space (L in Algorithm 1) is set to the same value as the number of categories for each dataset. To ensure a fair comparison for RONO [6] and HOPE [8], we replaced inputs without altering the core models and loss functions. Our FUME is implemented in PyTorch and trained on an Nvidia GTX 2080Ti GPU with a maximum of 200 epochs, using the Adam optimizer [48] with a learning rate of 5×10^{-5} for training.

The cosine similarity metric is exploited to measure similarity scores between different samples: $s(\mathbf{z}^i, \mathbf{z}^j) = (\mathbf{z}^i \cdot \mathbf{z}^j)/(|\mathbf{z}^i| \cdot |\mathbf{z}^j|)$. In the experiments, cross-modal retrieval performance is evaluated using the Mean Average Precision over all retrieved results (mAP@all) [49], which assesses both the precision and ranking quality of the results. It is computed by averaging the average precision (AP) across all retrieved results for each query.

4.2. Comparison With State-of-The-Art Methods

To verify the effectiveness of our FUME in cross-modal retrieval, we compare it with 13 state-of-the-art methods, including MCCA [22], ACMR [50], DSCMR [1], SDML [18], MAN [51], DRSL [52], ALGCN [53], ELRCMR [54], MARS [55], GNN4CMR [5], RONO [6], SCL [28] and HOPE [8]. For a fair comparison, all methods use the same image and text features with the default parameters provided by their authors. Notably, for the semi-supervised baselines (i.e., SCL [28] and HOPE [8]), their supervised variants are used for fair comparison, namely training with their supervised learning components only.

Experimental results are reported in Table 1. Here, 'FUME' does not employ our proposed uncertainty estimation method, while '$\text{FUME}_{u=0.5}$' incorporates this method with an uncertainty threshold set at 0.5. This threshold excludes retrieved results whose cross-modal uncertainty exceeds 0.5. The experimental results lead to the following observations: (1) Supervised methods, including our FUME, significantly outperform the unsupervised method (MCCA) because supervised models can leverage label information to learn more discriminative representations, thereby enhancing retrieval accuracy. (2) Our FUME method outperformed all baselines across the five datasets, achieving a notable 7.2% improvement on XMediaNet. These results highlight FUME's effectiveness for CMR, as it fully utilizes label information to optimize category credibility, learning more discriminative representations. (3) When our uncertainty estimation is employed to assist the model in retrieval, the retrieval performance of the model is significantly improved. Notably, upon setting the uncertainty threshold to 0.5, a performance boost is observed across all five datasets (16.9% on Pascal Sentence, 21.3% on Wikipedia, 18.0% on NUS-WIDE-10K, 32.7% on INRIA-Websearch, and 21.3% on XMediaNet). This underscores the effectiveness of the proposed cross-modal uncertainty in enhancing the retrieval capabilities of the model, thereby facilitating reliable and trusted retrieval outcomes. Additional comparisons, including the Precision-Recall curve, are available in the supplementary material.

4.3. Uncertainty Effect Analysis

To empirically validate the efficacy of our proposed cross-modal uncertainty estimation approach, we conduct a comprehensive comparative analysis against state-of-the-art methods. On the one hand, to show the effectiveness of cross-modal uncertainty over the similarity score, FUME is compared with the best two similarity-based baselines on each dataset. On the other hand, to further verify the advantage of the proposed uncertainty estimation method over the Evidential Deep Learning (EDL) [11] in CMR, FUME is also compared with the method that replaces Fuzzy Multimodal Learning in Section 3.2 with EDL and removes normalization for representations in the common space. To qualify the uncertainty of each retrieved result, for this method, Equation (10) is also employed to combine uncertainties across modalities for fair comparison. To compare these methods on the same metric, we introduce a metric:

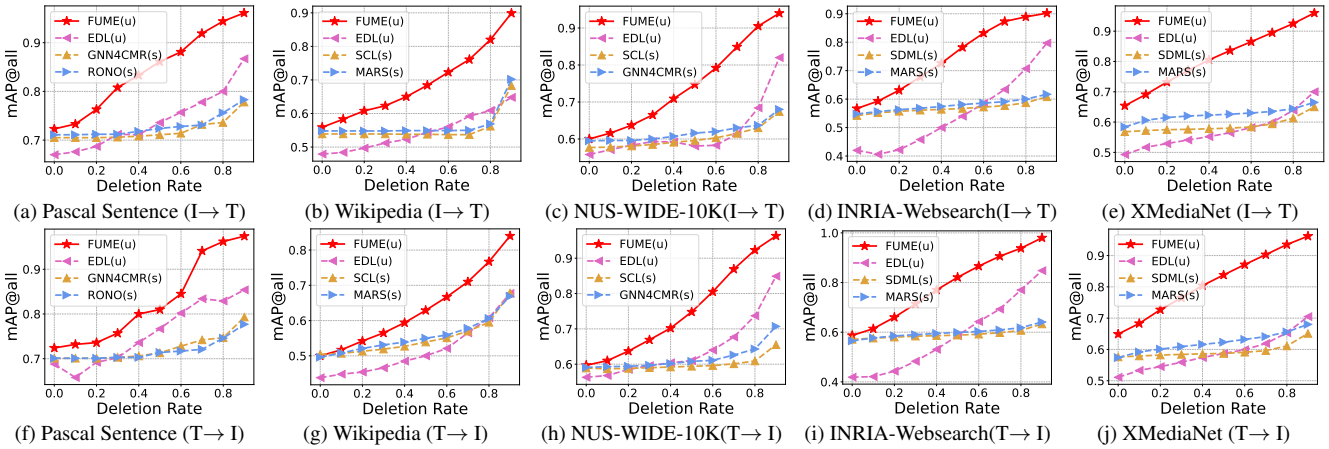
Deletion Rate (DR) is defined as the ratio of the number of retrieved samples removed based on specific criteria (e.g., similarity score or uncertainty) to the total number of original samples:

$$\text{DR} = \frac{\sum_{i=1}^{N_q}\sum_{j=1}^{N_d}\mathbb{I}_{ij}}{N_q \cdot N_d}, \quad \mathbb{I}_{ij} = \begin{cases} 1; & D_{ij} \geqslant t \\ 0; & D_{ij} < t \end{cases}, \tag{11}$$

where N_q is the number of query samples and N_d is the number of samples in the retrieval database, $D \in \mathbb{R}^{N_q \times N_d}$

Table 1. Performance comparison of mAP@all scores on Pascal Sentence, Wikipedia, NUS-WIDE-10K, INRIA-Websearch, and XMediaNet datasets. **Bold** font and underlined font indicate the highest and second-highest scores, respectively. The abbreviations of 'I' and 'T' represent Image and Text, respectively.

Methods	Ref.	Pascal Sentence			Wikipedia			NUS-WIDE-10K			INRIA-Websearch			XMediaNet		
		I→T	T→I	Avg.	I→T	T→I	Avg.	I→T	T→I	Avg.	I→T	T→I	Avg.	I→T	T→I	Avg.
MCCA [22]	SiKDD'2010	0.658	0.640	0.649	0.296	0.283	0.290	0.387	0.392	0.390	0.369	0.393	0.381	0.090	0.098	0.094
ACMR [50]	ACM MM'2017	0.663	0.591	0.627	0.434	0.414	0.424	0.537	0.542	0.540	0.420	0.426	0.423	0.506	0.466	0.486
DSCMR [1]	CVPR'2019	0.679	0.683	0.681	0.527	0.481	0.504	0.575	0.585	0.580	0.542	0.567	0.555	0.513	0.501	0.507
MAN [51]	KBS'2019	0.680	0.700	0.690	0.528	0.480	0.504	0.565	0.575	0.570	0.537	0.550	0.544	0.439	0.423	0.455
SDML [18]	SIGIR'2019	0.672	0.695	0.684	0.522	0.488	0.505	0.597	0.587	0.583	0.542	0.571	0.557	0.561	0.576	0.567
DRSL [52]	INS'2021	0.689	0.694	0.691	0.498	0.472	0.485	0.557	0.556	0.557	0.479	0.509	0.494	0.143	0.135	0.139
ALGCN [53]	TMM'2021	0.666	0.682	0.674	0.485	0.451	0.468	0.569	0.570	0.570	0.415	0.411	0.413	0.362	0.368	0.365
ELRCMR [54]	ACM MM'2022	0.705	0.700	0.702	0.543	0.501	0.521	0.549	0.567	0.558	0.300	0.283	0.292	0.056	0.073	0.065
MARS [55]	TCSVT'2022	0.678	0.685	0.682	0.548	0.496	0.522	0.551	0.555	0.553	0.545	0.564	0.555	0.584	0.574	0.579
GNN4CMR [5]	TPAMI'2023	0.705	0.701	0.703	0.521	0.480	0.501	0.595	0.591	0.593	0.523	0.538	0.531	0.566	0.568	0.567
RONO [6]	CVPR'2023	0.711	0.701	0.706	0.521	0.473	0.497	0.559	0.590	0.575	0.456	0.464	0.460	0.172	0.149	0.161
SCL [28]	TMM'2023	0.694	0.693	0.693	0.539	0.504	0.522	0.577	0.589	0.583	0.456	0.472	0.464	0.170	0.163	0.178
HOPE [8]	TPAMI'2024	0.672	0.691	0.681	0.499	0.474	0.487	0.538	0.570	0.554	0.480	0.504	0.492	0.516	0.515	0.515
FUME	Ours	0.723	0.724	0.723	0.558	0.501	0.530	0.600	0.597	0.598	0.567	0.587	0.577	0.654	0.649	0.651
FUME$_{\alpha=0.5}$	Ours	**0.890**	**0.894**	**0.892**	**0.767**	**0.719**	**0.743**	**0.775**	**0.781**	**0.778**	**0.870**	**0.938**	**0.904**	**0.860**	**0.868**	**0.864**



(a) Pascal Sentence (I→ T) (b) Wikipedia (I→ T) (c) NUS-WIDE-10K(I→ T) (d) INRIA-Websearch(I→ T) (e) XMediaNet (I→ T)

(f) Pascal Sentence (T→ I) (g) Wikipedia (T→ I) (h) NUS-WIDE-10K(T→ I) (i) INRIA-Websearch(T→ I) (j) XMediaNet (T→ I)

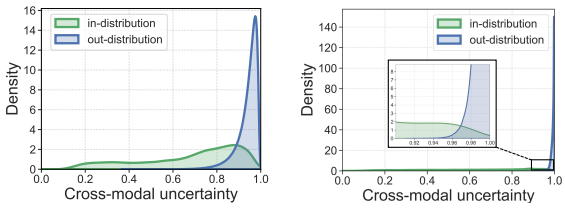
Figure 3. mAP@all across different deletion rates on five datasets (Pascal Sentence, Wikipedia, NUS-WIDE-10K, INRIA-Websearch, and XMediaNet). 'u' indicates that uncertainty is used to filter out retrieved results with high uncertainty, while 's' indicates that similarity score is used to filter out retrieved results with low similarity scores.

is the similarity matrix or cross-modal uncertainty matrix, t represents a threshold.

The experimental results are presented in Figure 3, from which we can make the following observations: (1) As the deletion rate increases, retrieval performance improves across all methods, but methods based on similarity score improve less due to ignoring the uncertainty of results and retaining implausible results post-filtering. (2) As the deletion rate increases, EDL gradually outperforms similarity-based methods across nearly all datasets, indicating that the uncertainty estimated by EDL can capture some incredible retrieved results that would degrade retrieval performance. However, EDL has limitations: Its performance is not as good as the most advanced methods at low deletion rates, and its retrieval performance declines with increasing deletion rates in Figure 3 (c), (d), and (f) due to imprecise uncertainty estimation. (3) FUME consistently outperforms all comparison methods at equal deletion rates, with its superiority becoming more pronounced as the deletion rate increases. This is due to the proposed FUME effectively capturing incredible retrieved results. Therefore, the results confirm FUME's ability to reduce performance losses from incredible results and also confirm FUME's ability to provide trusted predictions by providing precise uncertainty.

4.4. Identification of Out-of-distribution Data

To further evaluate the estimated cross-modal uncertainty, we visualize the distribution of in-/out-of-distribution samples in terms of cross-modal uncertainty in Figure 4. we could draw three observations: (1)The in-distribution and out-of-distribution curves rarely overlap, with out-of-distribution samples showing significantly higher uncertainties compared to in-distribution samples, thus confirming our method's ability to capture precise epistemic uncertainty. (2) In Figure 4 (b), the out-of-distribution uncertainty is higher than in Figure 4 (a) because XMediaNet,



(a) NUS-WIDE-10K (b) XMediaNet

Figure 4. Density of cross-modal uncertainty obtained by our FUME. For NUS-WIDE-10K, 'out-distribution' is the XMediaNet dataset, and for XMediaNet, it is the NUS-WIDE-10K dataset.

with 32,000 training samples, provides more sample for the neural network to learn its sample distribution than NUS-WIDE-10K, which has only 8,000 training samples. (3) The uncertainty density centers of in-distribution samples for both datasets are skewed to the right (i.e., greater than 0.5) because the neural network has not learned the features of all data within the distribution. Qualitative results can be found in the supplementary material.

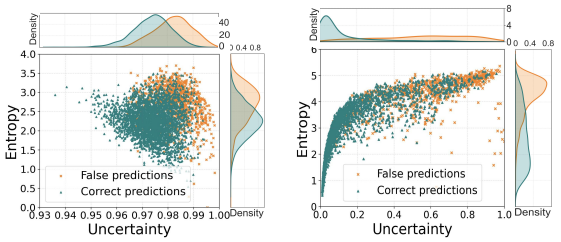
4.5. Counter-intuitive Problem Analysis

To investigate the counter-intuitive problem in EDL, we calculate the uncertainty and entropy of the classification results for EDL and our method on the image modality of the XMediaNet dataset's testing set, as shown in Figure 5. The observations are as follows: (1) On the whole, the entropy and uncertainty of correct predictions are lower than that of false predictions. (2) For EDL, the uncertainty is unrelated to the entropy, which implies that the uncertainty for conflicting and non-discriminatory evidence (high entropy) may be underestimated. In contrast, the decision uncertainty calculated by our method positively correlates with the entropy, thereby avoiding this problem. (3) EDL's uncertainty remains narrowly concentrated within the range of 0.93–1.00, despite a classification accuracy of 0.768, suggesting that EDL may overestimate uncertainty on large-category datasets such as the XMediaNet dataset with 200 categories. In contrast, our method mitigates this problem.

The root cause of the above counter-intuitive problems in EDL lies in its dependence on total evidence and category count for uncertainty estimation, overlooking how the evidence is distributed across each category. In contrast, our method achieves more precise uncertainty estimates by incorporating possibility and necessity measures, making it especially suitable for cross-modal retrieval applications. Implementation details for this experiment and more analysis are provided in the supplementary material.

4.6. Ablation Study

In this section, we conduct ablation experiments on the NUS-WIDE-10K and XMediaNet datasets to investigate the impact of individual components of the proposed FUME on retrieval performance. The results showcased in Table 2 lead to the following observations: (1) Both $\mathcal{L}_{fml}$ and $\mathcal{L}_{cl}$



(a) EDL (Accuracy = 0.768) (b) Ours (Accuracy=0.804)

Figure 5. Scatter plots of entropy and uncertainty of the classification results on the image modality in the XMediaNet testing set.

Table 2. Ablation studies on the NUS-WIDE-10K and XMediaNet datasets.

No.	$\mathcal{L}_{fml}$	$\mathcal{L}_{cl}$	NUS-WIDE-10K			XMediaNet		
			I→T	T→I	Avg.	I→T	T→I	Avg.
#1	✓		0.585	0.587	0.586	0.646	0.644	0.645
#2		✓	0.496	0.518	0.507	0.362	0.368	0.365
#3	✓	✓	**0.600**	**0.597**	**0.598**	**0.654**	**0.649**	**0.651**

are contribute to cross-modal retrieval in our framework. (2) Compared with $\mathcal{L}_{cl}$, the proposed $\mathcal{L}_{fml}$ has a greater impact on retrieval performance, especially on the XMediaNet dataset with 200 categories.

5. Conclusion

In this paper, we propose a novel method called Fuzzy Multimodal Learning (FUME) to capture and leverage the epistemic uncertainty in cross-modal retrieval. Our method frames the outputs of the classifier as a set of membership degrees and introduces a category credibility metric that incorporates both possibility and necessity measures. To address the counter-intuitive issue of existing uncertainty learning methods, we redefine decision uncertainty using the category credibility. Additionally, a merging scheme is proposed to combine the uncertainty from two modalities to qualify cross-modal uncertainty. This uncertainty, combined with the similarity scores, enables trusted cross-modal retrieval. Experimental results on five benchmark datasets demonstrate that our FUME remarkably improves both retrieval performance and reliability.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2024YFB4710604; in part by NSFC under Grant 62472295 and 62372315; in part by Sichuan Science and Technology Planning Project under Grant 2024NSFTD0047, 2024YFHZ0089, 2024NSFTD0049 and 2024YFHZ0144; in part by System of Systems and Artificial Intelligence Laboratory pioneer fund grant; in part by the Fundamental Research Funds for the Central Universities under Grant CJ202403; and in part by Chengdu Science and Technology Project (Grant no. 2023-XT00-00004-GX).

References

[1] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10394–10403, 2019. 1, 2, 6, 7

[2] Peng Hu, Liangli Zhen, Xi Peng, Hongyuan Zhu, Jie Lin, Xu Wang, and Dezhong Peng. Deep supervised multi-view learning with graph priors. *IEEE Transactions on Image Processing*, 33:123–133, 2023. 1, 2

[3] Xinran Ma, Mouxing Yang, Yunfan Li, Peng Hu, Jiancheng Lv, and Xi Peng. Cross-modal retrieval with noisy correspondence via consistency refining and mining. *IEEE Transactions on Image Processing*, 2024. 1

[4] Lei Zhu, Tianshi Wang, Fengling Li, Jingjing Li, Zheng Zhang, and Heng Tao Shen. Cross-modal retrieval: A systematic review of methods and future directions. *arXiv preprint arXiv:2308.14263*, 2023. 1

[5] Shengsheng Qian, Dizhan Xue, Quan Fang, and Changsheng Xu. Integrating multi-label contrastive learning with dual adversarial graph neural networks for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4794–4811, 2023. 1, 6, 7

[6] Yanglin Feng, Hongyuan Zhu, Dezhong Peng, Xi Peng, and Peng Hu. Rono: Robust discriminative learning with noisy labels for 2d-3d cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11610–11619, 2023. 2, 6, 7

[7] Yingying Huang, Bingliang Hu, Yipeng Zhang, Chi Gao, and Quan Wang. A semi-supervised cross-modal memory bank for cross-modal retrieval. *Neurocomputing*, page 127430, 2024.

[8] Fan Zhang, Hang Zhou, Xian-Sheng Hua, Chong Chen, and Xiao Luo. Hope: A hierarchical perspective for semi-supervised 2d-3d cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 6, 7

[9] Ruitao Pu, Yang Qin, Dezhong Peng, Xiaomin Song, and Huiming Zheng. Deep reversible consistency learning for cross-modal retrieval. *IEEE Transactions on Multimedia*, 2025. 1

[10] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk*, pages 139–147, 2010. 1, 6

[11] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018. 1, 2, 3, 6

[12] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023. 2

[13] Hao Wang and Dit-Yan Yeung. A survey on bayesian deep learning. *ACM computing surveys (csur)*, 53(5):1–37, 2020. 2

[14] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 2, 3

[15] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 2, 3

[16] Lotfi Asker Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965. 2, 3

[17] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Multi-view discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):188–194, 2015. 2

[18] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 635–644, 2019. 2, 6, 7

[19] Peng Hu, Xi Peng, Hongyuan Zhu, Liangli Zhen, and Jie Lin. Learning cross-modal retrieval with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5403–5413, 2021. 2, 4

[20] Yuan Sun, Jian Dai, Zhenwen Ren, Yingke Chen, Dezhong Peng, and Peng Hu. Dual self-paced cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15184–15192, 2024. 2

[21] Bruce Thompson. *Canonical correlation analysis: Uses and interpretation*, volume 47. Sage, 1984. 2

[22] Jan Rupnik and John Shawe-Taylor. Multi-view canonical correlation analysis. In *Conference on data mining and data warehouses (SiKDD 2010)*, volume 473, pages 1–4, 2010. 6, 7

[23] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. Cross-modality correlation propagation for cross-media retrieval. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2337–2340. IEEE, 2012. 2

[24] Shupeng Su, Zhisheng Zhong, and Chao Zhang. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3027–3035, 2019. 2

[25] Yongxiang Li, Yang Qin, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Romo: Robust unsupervised multimodal learning with noisy pseudo labels. *IEEE Transactions on Image Processing*, 33:5086–5097, 2024.

[26] Jinrong Cui, Zhipeng He, Qiong Huang, Yulu Fu, Yuting Li, and Jie Wen. Structure-aware contrastive hashing for unsupervised cross-modal retrieval. *Neural Networks*, page 106211, 2024. 2

[27] Peng Hu, Hongyuan Zhu, Xi Peng, and Jie Lin. Semi-supervised multi-modal learning with balanced spectral decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 99–106, Apr. 2020. 2

[28] Yaxin Liu, Jianlong Wu, Leigang Qu, Tian Gan, Jianhua Yin, and Liqiang Nie. Self-supervised correlation learning

for cross-modal retrieval. *IEEE Transactions on Multimedia*, 25:2851–2863, 2023. 6, 7

[29] Fan Zhang, Xian-Sheng Hua, Chong Chen, and Xiao Luo. Fine-grained prototypical voting with heterogeneous mixup for semi-supervised 2d-3d cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17016–17026, 2024. 2

[30] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4948–4956, 2022. 2

[31] Peng Hu, Xi Peng, Hongyuan Zhu, Jie Lin, Liangli Zhen, and Dezhong Peng. Joint versus independent multiview hashing for cross-view retrieval. *IEEE Transactions on Cybernetics*, 51(10):4982–4993, 2020. 2

[32] Peng Hu, Hongyuan Zhu, Jie Lin, Dezhong Peng, Yin-Ping Zhao, and Xi Peng. Unsupervised contrastive cross-modal hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3877–3889, 2022.

[33] Yuan Sun, Kaiming Liu, Yongxiang Li, Zhenwen Ren, Jian Dai, and Dezhong Peng. Distribution consistency guided hashing for cross-modal retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5623–5632, 2024.

[34] Kaiming Liu, Yunhong Gong, Yu Cao, Zhenwen Ren, Dezhong Peng, and Yuan Sun. Dual semantic fusion hashing for multi-label cross-modal retrieval. In *International Joint Conferences on Artificial Intelligence Organization, IJCAI*, pages 4569–4577, 2024. 2

[35] Baoding Liu and Baoding Liu. *Uncertainty theory*. Springer, 2010. 3

[36] Rangan Das, Sagnik Sen, and Ujjwal Maulik. A survey on fuzzy deep neural networks. *ACM Computing Surveys (CSUR)*, 53(3):1–25, 2020. 3

[37] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 4

[38] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 4

[39] Aldo De Luca and Settimo Termini. A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. In *Readings in fuzzy sets for intelligent systems*, pages 197–202. Elsevier, 1993. 5

[40] Pingke Li and Baoding Liu. Entropy of credibility distributions for fuzzy variables. *IEEE Transactions on Fuzzy Systems*, 16(1):123–129, 2008. 5

[41] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R.G. Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 251–260, New York, NY, USA, 2010. Association for Computing Machinery. 6

[42] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009. 6

[43] Josip Krapac, Moray Allan, Jakob Verbeek, and Frédéric Juried. Improving web image search results using query-relative classifiers. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1094–1101. IEEE, 2010. 6

[44] Yuxin Peng, Jinwei Qi, and Yuxin Yuan. Modality-specific cross-modal similarity measurement with recurrent attention network. *IEEE Transactions on Image Processing*, 27(11):5585–5599, 2018. 6

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[46] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 6

[47] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 6

[48] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[49] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. Generalized multiview analysis: A discriminative latent space. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2160–2167. IEEE, 2012. 6

[50] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 154–162, 2017. 6, 7

[51] Peng Hu, Dezhong Peng, Xu Wang, and Yong Xiang. Multimodal adversarial network for cross-modal retrieval. *Knowledge-Based Systems*, 180:38–50, 2019. 6, 7

[52] Xu Wang, Peng Hu, Liangli Zhen, and Dezhong Peng. Drsl: Deep relational similarity learning for cross-modal retrieval. *Information Sciences*, 546:298–311, 2021. 6, 7

[53] Shengsheng Qian, Dizhan Xue, Quan Fang, and Changsheng Xu. Adaptive label-aware graph convolutional networks for cross-modal retrieval. *IEEE Transactions on Multimedia*, 24:3520–3532, 2021. 6, 7

[54] Tianyuan Xu, Xueliang Liu, Zhen Huang, Dan Guo, Richang Hong, and Meng Wang. Early-learning regularized contrastive learning for cross-modal retrieval with noisy labels. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 629–637, New York, NY, USA, 2022. Association for Computing Machinery. 6, 7

[55] Yunbo Wang and Yuxin Peng. Mars: Learning modality-agnostic representation for scalable cross-media retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4765–4777, 2021. 6, 7