Fuzzy Multimodal Learning for Trusted Cross-modal Retrieval

Supplementary Material

In this supplementary material, we provide complementary information on theory and experiments. Specifically, in Appendix A, we supplement the proofs of mathematical properties of decision uncertainty and cross-modal uncertainty. In Appendix B, we offer additional experimental settings and experimental results. More specifically, Appendix B.1 introduces a detailed overview of the datasets used in all experiments, Appendix B.2 provides a comprehensive description of the baselines, Appendix B.3 presents precision-recall curve comparisons, Appendix B.4 visualizes the effectiveness of cross-modal uncertainty, Appendix B.5 provides the qualitative results for OOD detection, Appendix B.6 provides visualization of the learned representation by applying t-SNE [1] method, Appendix B.7 offers additional counter-intuitive problem analysis, Appendix B.8 provide parameter analysis, and Appendix B.9 compares different uncertainty merging schemes and proves the superiority of the proposed scheme.

A. Mathematical Proof of Properties

A.1. Proof of the Properties for Decision Uncertainty

In this section, we further supplement **Section 3.5** with properties of decision uncertainty and corresponding clear proof. First, we begin by reviewing the 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 the i-th sample of j-th modality and $\forall c_{ik}^j \in [0,1],\ k = 1, 2, ..., K$. Then, the decision uncertainty is defined by

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

where $H(\mathbf{c}_i^j)$ is the entropy of category credibility, K is the number of categories, and $S(t) = -t \ln t - (1-t)\ln(1-t)$. This decision uncertainty is in the range $[0, 1]$, and has the following properties:

Property 1 (Lower bound): 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 the i-th sample of j-th modality and $\forall c_{ik}^j \in [0,1],\ k = 1, 2, ..., K$. Therefore,

$$u_i^j = U(\mathbf{c}_i^j) \geqslant 0, \tag{2}$$

and the equality holds if and only if $\forall c_{ik}^j \in \{0, 1\}$.

Proof 1: Since function $S(t)$ reaches its minimum 0 at $t = 0$ or $t = 1$, we have

$$H(\mathbf{c}_i^j) = \sum_{k=1}^K S(c_{ik}^j) \geqslant 0, \tag{3}$$

and then

$$u_i^j = U(\mathbf{c}_i^j) = \frac{H(\mathbf{c}_i^j)}{K \cdot \ln 2} \geqslant 0, \tag{4}$$

the equation holds if and only if $\forall c_{ik}^j \in \{0, 1\}, k = 1, 2, ..., K$. This completes the proof.

Property 2 (upper bound): 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 the i-th sample of j-th modality and $\forall c_{ik}^j \in [0,1],\ k = 1, 2, ..., K$. Therefore, the uncertainty

$$u_i^j = U(\mathbf{c}_i^j) \leqslant 1 \tag{5}$$

and the equality holds if and only if $\forall c_{ik}^j = 0.5,\ k = 1, 2, ..., K$.

Proof 2: Since the function $S(t)$ reaches its maximum $\ln 2$ at $t = 0.5$, e.g, $S(t) \leqslant \ln 2$. Therefore, we have

$$H(\mathbf{c}_i^j) = \sum_{k=1}^K S(c_{ik}^j) \leqslant K \cdot \ln 2 \tag{6}$$

and then

$$u_i^j = U(\mathbf{c}_i^j) = \frac{H(\mathbf{c}_i^j)}{K \cdot \ln 2} \leqslant 1, \tag{7}$$

the equation holds if and only if $\forall c_{ik}^j = 0.5, k = 1, 2, ..., K$. This completes the proof.

Property 3 (Symmetry): Let $\mathbf{c}_1^j = [c_{11}^j, c_{12}^j, ..., c_{1K}^j]$ be the vector of credibility degrees, and let $\mathbf{c}_2^j = [c_{21}^j, c_{22}^j, ..., c_{2K}^j]$ be the another vector of credibility degrees. If $[c_{11}^j, c_{12}^j, ..., c_{1K}^j]$ is a rearrangement of $[c_{21}^j, c_{22}^j, ..., c_{2K}^j]$, Then we have $U(\mathbf{c}_1^j) = U(\mathbf{c}_2^j)$.

Proof 3: The property follows immediately from the definition of entropy. The symmetry property establishes that decision uncertainty remains consistent regardless of the permutations of credibility degrees.

A.2. Proof of the Properties for Cross-modal Uncertainty

In this section, we provide supplementary material for **Section 3.5**, including proofs that demonstrate how cross-modal uncertainty satisfies the four specified properties. We begin by revisiting the definition of cross-modal uncertainty:

Table 1. General statistics of the datasets used in the experiments, where '* / * / *' represent the respective counts of training, validation, and testing image-text pairs. The symbol K denotes the total number of categories, while d_I and d_T signify the dimensionalities of the image and text features obtained by VGGNet [2] and word2vec [3], respectively.

Dataset	train / val / test	K	d_I	d_T
Pascal Sentence [4]	800 / 100 / 100	20	4,096	300
Wikipedia [5]	2,173 / 231 / 462	10	4,096	300
NUS-WIDE-10K [6]	8,000 / 1,000 / 1,000	10	4,096	300
INRIA-Websearch [7]	9,000 / 1,332 / 4,366	100	4,096	1,000
XMediaNet [8]	32,000 / 4,000 / 4,000	200	4,096	300

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

$$u_i^{1\Leftrightarrow 2} = g(u_i^1, u_i^2) = 1 - (1-u_i^1)(1-u_i^2). \tag{8}$$

Its properties and the corresponding mathematical proof are as follows:

Property 1: $0 \leqslant u_i^{1\Leftrightarrow 2} \leqslant 1$.

Proof 1: Because $u_i^1 \in [0,1]$ and $u_i^2 \in [0,1]$, it's pretty obvious that

$$\begin{aligned} 1 &\geqslant (1-u_i^1)(1-u_i^2) \geqslant 0 \\ 1 &\geqslant 1-(1-u_i^1)(1-u_i^2) \geqslant 0 \end{aligned} \tag{9}$$

where the left equal sign is true if and only if u_i^1 and $u_i^2 = 0$, and the right equal sign is true if and only if u_i^1 and $u_i^2 = 1$. This completes the proof.

Property 2: $u_i^{1\Leftrightarrow 2} \geqslant \max(u_i^1, u_i^2)$.

Proof 2: Assuming that

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

so we have

$$2u_i^{1\Leftrightarrow 2} = 2-2(1-u_i^1)(1-u_i^2) \geqslant u_i^1 + u_i^2, \tag{11}$$

it can be deduced as:

$$\begin{aligned} 2-2(1-u_i^1-u_i^2+u_i^1u_i^2) &\geqslant u_i^1+u_i^2 \\ u_i^1+u_i^2 &\geqslant 2u_i^1u_i^2. \end{aligned} \tag{12}$$

Therefore, if $u_i^1 + u_i^2 \geqslant 2u_i^1u_i^2$ is true, $1-(1-u_i^1)(1-u_i^2) \geqslant \max(u_i^1, u_i^2)$ is also true. So, we below prove that $u_i^1 + u_i^2 \geqslant 2u_i^1u_i^2$ is true. To be specific, given by AM–GM inequality:

$$u_i^1 + u_i^2 \geqslant 2\sqrt{u_i^1u_i^2}. \tag{13}$$

When $u_i^1 \in [0,1]$ and $u_i^2 \in [0,1]$ is true, it's obvious that

$$\sqrt{u_i^1u_i^2} \geqslant u_i^1u_i^2. \tag{14}$$

Combine Equation (13) and Equation (14), we can easily get

$$u_i^1 + u_i^2 \geqslant 2u_i^1u_i^2. \tag{15}$$

Therefore, the hypothesis that $u_i^{1\Leftrightarrow 2} \geqslant \max(u_i^1, u_i^2)$ is valid and this property is verified.

Property 3: if $u_i^1 = 1$ or $u_i^1 = 1$, $u_i^{1\Leftrightarrow 2} = 1$.

This does not require formal proof, as it is straightforward to deduce.

Property 4: if $u_2^1 \geqslant u_1^1$ and $u_2^2 \geqslant u_1^2$, $u_2^{1\Leftrightarrow 2} \geqslant u_1^{1\Leftrightarrow 2}$.

Proof 4: if $u_2^1 \geqslant u_1^1$ and $u_2^2 \geqslant u_1^2$, $u_2^{1\Leftrightarrow 2} \geqslant u_1^{1\Leftrightarrow 2}$: Assuming that $u_2^1 \geqslant u_1^1$ and $u_2^2 \geqslant u_1^2$, because $u_1^1,\ u_2^1,\ u_1^2$ and $u_2^2 \in [0,1]$, we can deduced that $1-u_2^1 \leqslant 1-u_1^1$ and $1-u_2^2 \leqslant 1-u_1^2$, and then

$$\begin{aligned} (1-u_2^1)(1-u_2^2) &\leqslant (1-u_1^1)(1-u_1^2) \\ 1-(1-u_2^1)(1-u_2^2) &\geqslant 1-(1-u_1^1)(1-u_1^2) \\ u_2^{1\Leftrightarrow 2} &\geqslant u_1^{1\Leftrightarrow 2}. \end{aligned} \tag{16}$$

This completes the proof.

B. Additional Experiments

Some experimental details and results could not be fully presented in the main paper due to space constraints. This section complements additional details, results, and analyses.

B.1. Datasets

Without compromising generality, we utilize five commonly utilized image-text datasets to assess the cross-modal performance in this paper. The details about those experimental datasets are listed as follows, and the statistical results of the datasets are summarized in Table 1.

- **Pascal Sentence** [4] comprises 1000 image-text pairs. Images are sourced from the 2008 PASCAL development kit, while each corresponding text sample contains five independent sentences, annotated by different individuals on Amazon Mechanical Turk. For a fair comparison, the image-text pairs of the dataset are randomly selected to constitute 3 sets: the training set with 800 image-text

pairs, the validation set with 100 image-text pairs, and the testing set with 100 image-text pairs.

- **Wikipedia** [5] is a widely-used dataset for cross-media retrieval. It contains $2,866$ image-text pairs and each image or text sample is classified into 10 classes (*i.e.*, literature, media, music, etc.). In the experiment, we sample $2,173$ pairs as the training set, 231 as the validation set, and 462 as the testing set.
- **NUS-WIDE-10K** [6] utilized in this study is a subset of the NUS-WIDE dataset [9] provided by the authors of Reference [6]. The author selected an equal number of image-text pairs (1000 pairs per class) from the top ten most populous categories (such as grass, person, sky, etc.) within the NUS-WIDE dataset, creating the NUS-WIDE-10K subset. In the experiment, we sample $8,000$ pairs as the training set, $2,000$ as the validation set, and $2,000$ as the testing set.
- **INRIA-Websearch** [7] originally comprises 71,478 images, each accompanied by a corresponding text description. In this study, we utilize a specific subset of it, curated by the authors of [10]. This subset focuses on 14,698 samples, drawn from the 100 most populous classes in the original collection. We randomly divide the dataset into three subsets: 9,000, 1,332, and 4,366 image-text pairs for training, validation, and testing sets, respectively.
- **XMediaNet** [8] is a large-scale multimodal dataset, which comprises 5 media types, *i.e.*, image, text, video, audio, and 3-D model. It comprises 40,000 images, 40,000 texts, 10,000 videos, 10,000 audio clips, and 2,000 3D models from 200 semantic classes. Each sample of different modalities has 200 non-overlap categories, For this study, we only focus on images and text. In our experiments, we sample 32,000 pairs for the training set, 4,000 for the validation set, and 4,000 for the testing set.

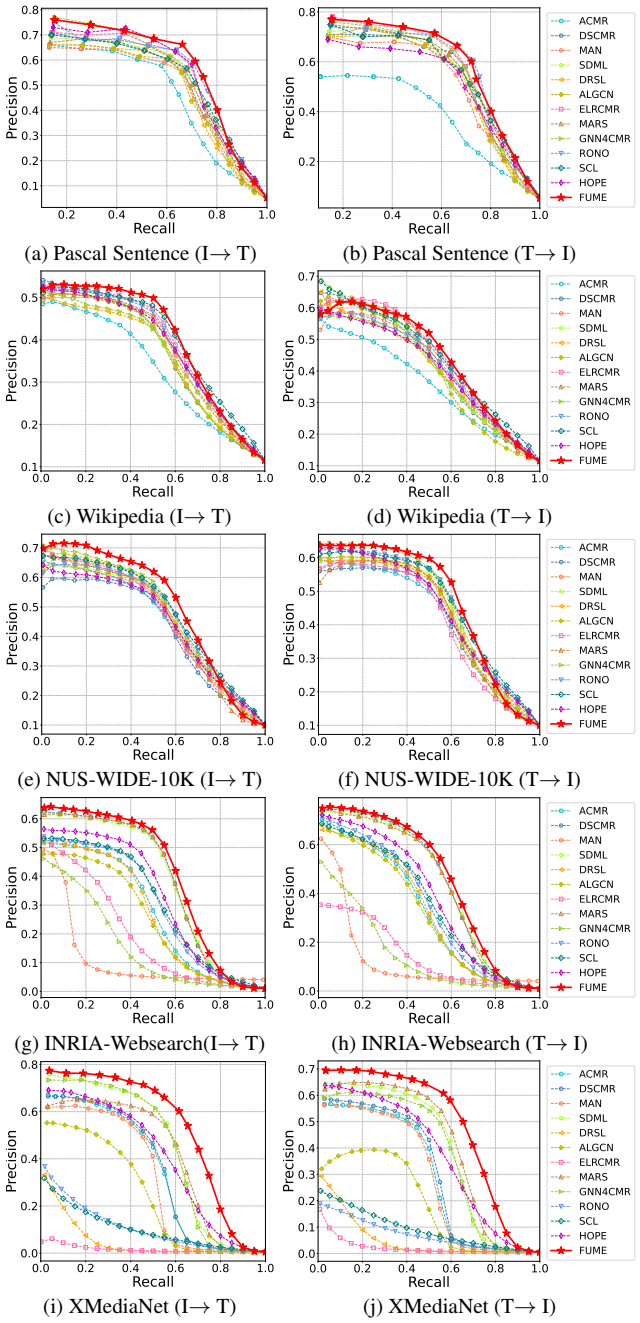
B.2. Baselines

To verify the effectiveness of the proposed method in cross-modal retrieval, we compare 13 state-of-the-art methods in the experiments, including MCCA [11], ACMR [12], DSCMR [13], SDML [14], MAN [15], DRSL [16], ALGCN [17], ELRCMR [18], MARS [19], GNN4CMR [20], RONO [21], SCL [22] and HOPE [23]. We briefly introduce these compared methods in the following paragraphs:

- **MCCA** [11] learns multiple modality-specific linear transformations to map the different modalities into a common space by maximizing the correlations between all possible pairwise modalities.
- **ACMR** [12] utilizes adversarial learning to discover an effective common subspace, incorporating a feature projector, a modality classifier, and a triplet constraint.
- **DSCMR** [13] is an early exploration into deep cross-modal retrieval, minimizing both discrimination loss and modality invariance loss to derive shared representations across diverse modalities.
- **SDML** [14] is one of the first works to independently project data of an unfixed number of modalities into a predefined common subspace.
- **MAN** [15] pioneers multimodal representation learning (involving more than two modalities) through adversarial learning.
- **DRSL** [16] is the first approach that incorporates relation networks into cross-modal learning, effectively bridging the heterogeneity gap between different modalities by directly learning natural pairwise similarities.
- **ALGCN** [17] uncovers the semantics of labels and conserves semantic correlations across different modalities through the joint training of two different branches.
- **ELRCMR** [18] solves the problem of noise label memorization and cluster drift in cross-modal retrieval by employing early learning regularization contrast learning and dynamic weight balance strategy.
- **MARS** [19] treat label information as a distinct modality and realize scalable cross-modal retrieval, which allows each modality to be trained independently.
- **GNN4CMR** [20] integrates multi-label contrastive learning with dual adversarial graph neural networks for cross-modal retrieval.
- **RONO** [21] tackles the problem of 2D-3D retrieval under label noise by proposing a consistency loss and a robust center learning strategy.
- **SCL** [22] uses unlabeled data to learn discriminative and unsupervised contrast learning to model intra-modal and inter-modal instance relationships to improve retrieval performance.
- **HOPE** [23] is a semi-supervised cross-modal retrieval method that effectively deals with the complex relationship between 2D and 3D data through hierarchical alignment and fuzzy pseudo-label technology.

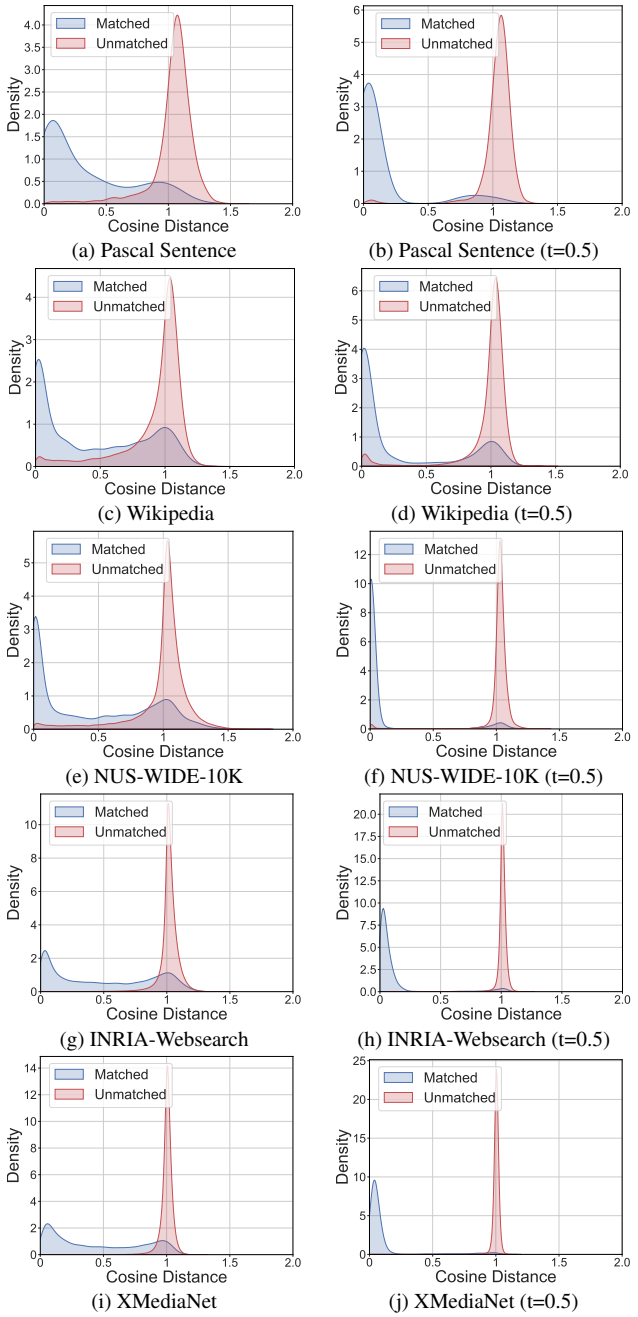
B.3. Precision-recall Curve Comparisons

To further comprehensively compare the retrieval performance between our FUME and the other baselines (i.e., ACMR [12], DSCMR [13], SDML [14], MAN [15], DRSL [16], ALGCN [17], ELRCMR [18], MARS [19], GNN4CMR [20], RONO [21], SCL [22] and HOPE [23]), we plotted Precision-recall (P-R) curves as shown in Figure 1. Precision and recall represent two conflicting metrics: an increase in one often leads to a decrease in the other. In P-R curve plots, methods represented by curves positioned higher on the graph typically demonstrate better performance. The results show that our FUME outperforms the other baselines in both image-query-text and text-query-image retrieval tasks, especially on large datasets. This further confirms the effectiveness of our FUME in cross-modal retrieval.



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

Figure 1. Precision-recall curves for the image-query-texts (I → T) and text-query-images (T → I) experiments on the Pascal Sentence, Wikipedia, NUS-WIDE-10K, INRIA-Websearch, and XMediaNet datasets. The abbreviations of 'I' and 'T' represent image and text respectively.



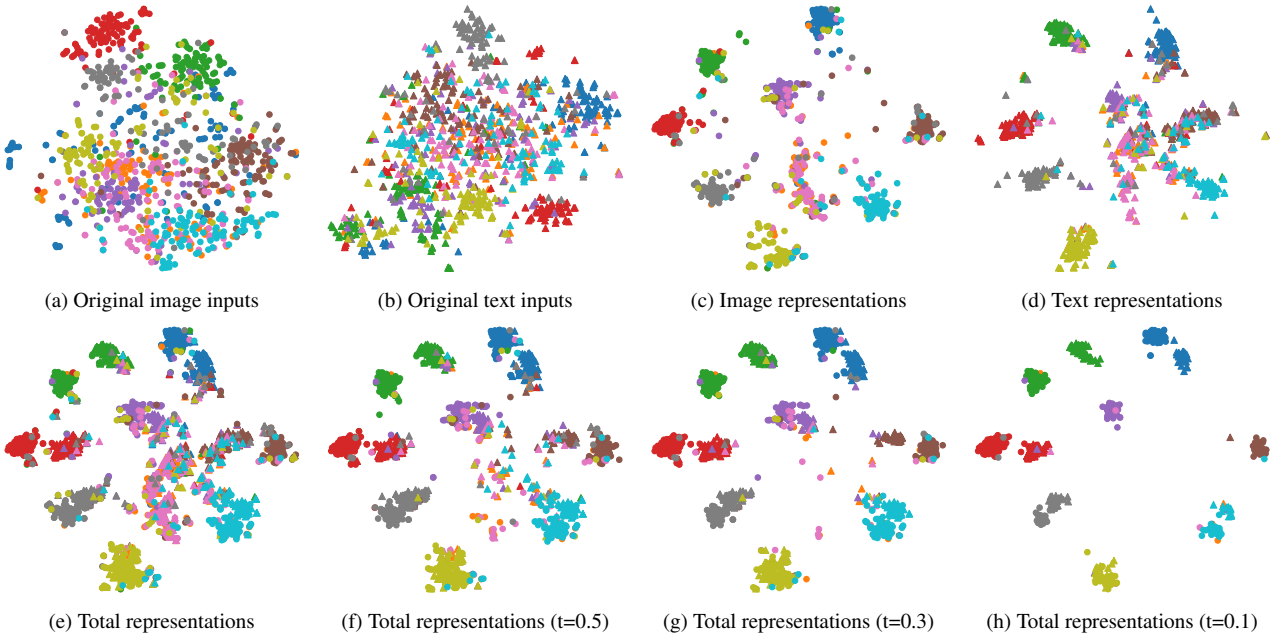
(a) Pascal Sentence (b) Pascal Sentence (t=0.5)
(c) Wikipedia (d) Wikipedia (t=0.5)
(e) NUS-WIDE-10K (f) NUS-WIDE-10K (t=0.5)
(g) INRIA-Websearch (h) INRIA-Websearch (t=0.5)
(i) XMediaNet (j) XMediaNet (t=0.5)

Figure 2. Density of cosine distance of representations in the common space of on the Pascal Sentence, Wikipedia, NUS-WIDE-10K, INRIA-Websearch, and XMediaNet datasets. 't=0.5' means setting the cross-modal uncertainty threshold to 0.5, i.e., the samples whose cross-modal uncertainty greater than 0.5 are removed.

B.4. Visualization of Cross-Modal Uncertainty Effectiveness

To further evaluate the proposed FUME in capturing incredible retrieved results, we visualize the cosine distance (a lower cosine distance indicates higher similarity) density of representations in the common space. Results are shown in Figure 2, from which the following observation can be drawn:

- In Figure 2 (a), (c), (e), (g), and (i), it is challenging to dis-



(a) Original image inputs (b) Original text inputs (c) Image representations (d) Text representations

(e) Total representations (f) Total representations (t=0.5) (g) Total representations (t=0.3) (h) Total representations (t=0.1)

Figure 3. Visualizing the representations of image and text data in the testing set of the NUS-WIDE-10K [6] dataset by *t*-SNE [1]. Circles represent samples originating from the image modality, while triangles correspond to samples from the text modality. Samples belonging to the same semantic category are consistently marked with the same color. (a) Original image representation. (b) Original text representation. (c) Image representation on the common representation space. (d) Text representation on the common representation space. (e) Image and text representations (total representations) in the common space. (f)-(h) Image and text representations (total representations) in the common space after the sample with decision uncertainty greater than 0.5, 0.3, and 0.1 are removed, respectively.

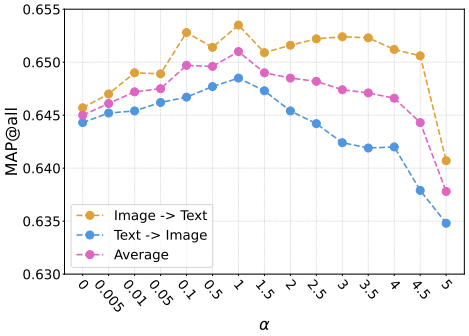
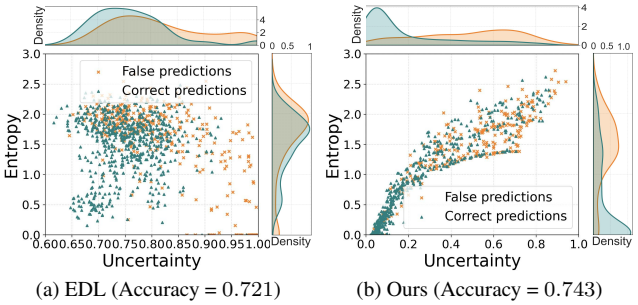


Figure 4. FUME's cross-modal retrieval performance versus different values of α on the testing set of the XMediaNet datasets.



(a) EDL (Accuracy = 0.721) (b) Ours (Accuracy = 0.743)

Figure 5. Scatter plots of entropy and uncertainty of the classification results on the image modality in the testing set of NUS-WIDE-10K dataset.

tinguish between matched and unmatched samples solely based on the cosine distance.

- After removing retrieved results whose cross-modal uncertainty is greater than 0.5, the matched and unmatched overlap significantly less, as shown in Figure 2 (b), (d), (f), (h), and (j). This means the proposed cross-modal uncertainty precisely captures incredible results and can thus be used to improve retrieval performance and achieve trusted cross-modal retrieval.

B.5. Qualitative results for OOD detection

To supplement **Section 4.4**, we conduct quantitative experiments for out-of-distribution (OOD) detection, using the False Positive Rate at 95% True Positive Rate (FPR95) as the evaluation metric. : When using NUS-WIDE-10K as in-distribution (ID) data and XMediaNet as OOD data, our FUME achieves an FPR95 of 0.4, remarkably lower than EDL's 0.959. Conversely, when XMediaNet is used as ID data and NUS-WIDE-10K as OOD data, FUME achieves an FPR95 of 0.002, better than EDL's 0.990. Both qualitative

and quantitative results consistently demonstrate FUME's superiority in OOD detection.

B.6. Visualisation of the Learned Representation

To visually investigate the effectiveness of the proposed FUME, we adopt the *t*-SNE [1] method to embed the representations of the image and text samples into a two-dimensional visualization plane using the NUS-WIDE-10K [6] dataset. The results of the original images represented by the 4,096-dimensional (VGGNet [2]) features and the text samples represented by the 300-dimensional (word2vec [3]) features (after the embedding process) are displayed in Figure 3 (a) and Figure 3 (b), respectively. The results show significant distribution differences between the image and text modalities in the NUS-WIDE-10K dataset, highlighting the challenges of sample classification in the original input space.

Figure 3 (c) and Figure 3(d) give the distribution of the learned representation of image and text, respectively. From these results, the proposed FUME contributes to discriminating the samples with different semantic categories, and several clusters show discriminative intervals. Figure 3 (e) demonstrates the distributions of image modality and text modality representations. From the results, we can see that the distributions of the image and text modalities exhibit considerable overlap, rendering them challenging to discern. This overlap underscores the significant reduction in cross-modal discrepancy achieved through the employment of our proposed FUME method.

To further estimate the effectiveness of decision uncertainty, we remove samples based on their decision uncertainty. As shown in Figure 3 (f)-(h), after the samples with decision uncertainty greater than 0.5, 0.3, and 0.1 were filtered out successively, the overlap of different categories was alleviated, and the boundaries of different categories became clearer. This improvement is thanks to the effective uncertainty estimation capability of the proposed method.

B.7. Additional Counter-intuitive Problem Analysis

To further investigate the counter-intuitive problem in Evidence Deep Learning (EDL) [24], we calculate the uncertainty and entropy of the classification results for EDL and our method on the image modality of the NUS-WIDE-10K and XMediaNet datasets' testing set. Specifically, for a fair comparison, we use the method that replaces Fuzzy Multimodal Learning with EDL and removes normalization for representations in the common space to output the evidence in the range $[0, +\infty]$.

B.7.1. Evaluation Metric

To evaluate the conflict of the prediction, we introduce the entropy (H) of the classification results:

$$H(\mathbf{p}_i^j) = \sum_{k=1}^{K} -p_{ik}^j \cdot \ln\left(p_{ik}^j\right) - (1-p_{ik}^j) \cdot \ln\left(1-p_{ik}^j\right), \quad (17)$$

where $\mathbf{p}_i^j = [p_{i1}^j, p_{i2}^j, ..., p_{ik}^j]$, $\sum_{k=1}^{K} p_{ik}^j = 1$, and p_{ik}^j represents classification probability of i-th sample of the j-th modality for k-th category. The more uniform the probability distribution, the greater the entropy, and it reaches its maximum value when the probability distribution is completely uniform, i.e., $\forall p_{ik}^j = \frac{1}{K}$. Therefore, the entropy is a conflicting evaluation metric that measures the conflict of the prediction, higher entropy reflects higher conflict. For a fair comparison, the classification probability p_{ik}^j in EDL is calculated by:

$$p_{ik}^j = \frac{e_{ik}^j}{\sum_{k=1}^{K} e_{ik}^j}, \quad (18)$$

where e_{ik}^j represents the evidence of the i-th sample of j-th modality for the k-th category. For our method, the p_{ik}^j is calculated by:

$$p_{ik}^j = \frac{m_{ik}^j}{\sum_{k=1}^{K} m_{ik}^j}, \quad (19)$$

where m_{ik}^j represents the membership degree of the i-th sample of j-th modality for the k-th category.

The uncertainty (u) of classification results in EDL is calculated by

$$u_i^j = \frac{K}{\sum_{k=1}^{K} e_{ik}^j + K}, \quad (20)$$

where K is the total number of categories. This formula is dependent on total evidence and category count for uncertainty estimation, overlooking how the evidence is distributed across each category.

B.7.2. Additional Analysis

We additionally compare EDL and our method on the NUS-WIDE-10K dataset, which has 10 categories (significantly fewer than the 200 categories in the XMediaNet dataset). The experimental results, shown in Figure 5, lead to the following observations:

- Similar to the results observed on the XMediaNet dataset in **Section 4.5**, EDL's uncertainty is unrelated to entropy. This suggests that uncertainty for conflicting and non-discriminatory evidence (high entropy) is likely underestimated. Conversely, the decision uncertainty estimated by our method positively correlates with the entropy, thereby achieving a more precise uncertainty estimation.

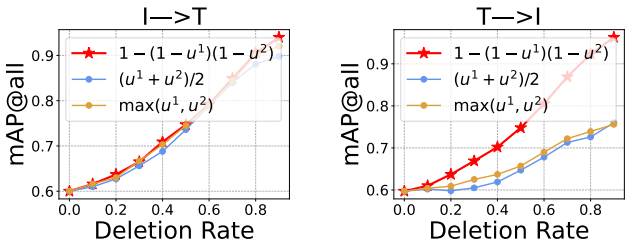


Figure 6. mAP@all across different deletion rates on the NUS-WIDE-10K datasets.

- EDL's uncertainty remains narrowly concentrated within the range of 0.60–1.00, even with a relatively high classification accuracy of 0.721. This indicates that the total evidence remains small compared to the number of categories, even with only 10 categories. Conversely, the uncertainty estimated by our method spans the full range [0, 1], indicating a more precise uncertainty estimation.

B.8. Parameter Analysis

To evaluate the influence of the trade-off hyper-parameter α in the loss function, we present a plot in Figure 4 illustrating the relationship between retrieval performance and α based on the testing set of XMediaNet. As shown, performance rises gradually from $\alpha = 0$, peaks between 0.005 and 2, and then declines. This trend underscores the role of $\mathcal{L}_{mcl}$ in enhancing multimodal discrimination, aligning with our ablation results in **Section 4.6**.

B.9. Uncertainty Merging Schemes Analysis

To verify this, we compared our merging scheme, $1-(1-u^1)(1-u^2)$, with the other two schemes $(u^1+u^2)/2$ and $\max(u^1,u^2)$ on the NUS-WIDE-10K dataset. The results, shown in the Figure 6, demonstrate that our scheme outperforms others, particularly achieving remarkable superiority in text-to-image retrieval.

References

[1] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 1, 5, 6

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 6

[3] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2, 6

[4] 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. 2

[5] 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. 2, 3

[6] 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. 2, 3, 5, 6

[7] 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. 2, 3

[8] 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. 2, 3

[9] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, page 7–16, New York, NY, USA, 2014. Association for Computing Machinery. 3

[10] Yunchao Wei, Yao Zhao, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, and Shuicheng Yan. Cross-modal retrieval with cnn visual features: A new baseline. *IEEE transactions on cybernetics*, 47(2):449–460, 2016. 3

[11] 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. 3

[12] 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. 3

[13] 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. 3

[14] 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. 3

[15] Peng Hu, Dezhong Peng, Xu Wang, and Yong Xiang. Multimodal adversarial network for cross-modal retrieval. *Knowledge-Based Systems*, 180:38–50, 2019. 3

[16] Xu Wang, Peng Hu, Liangli Zhen, and Dezhong Peng. Drsl: Deep relational similarity learning for cross-modal retrieval. *Information Sciences*, 546:298–311, 2021. 3

[17] 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. 3

[18] 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. 3

[19] 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. 3

[20] 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. 3

[21] 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. 3

[22] 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. 3

[23] 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. 3

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